

Bernard Schutz

Gravity from the ground up

CAMBRIDGE

An introductory guide to gravity and general relativity

CAMBRIDGE

www.cambridge.org/9780521455060

This page intentionally left blank

Gravity from the ground up

Gravity from the ground up

Bernard Schutz

*Max Planck Institute for
Gravitational Physics
(The Albert Einstein Institute)
Golm, Germany*

*and
Department of Physics and Astronomy
Cardiff University, UK*

CAMBRIDGE UNIVERSITY PRESS

Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

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/9780521455060

© Bernard Schutz 2003

This publication is in copyright. Subject to statutory exception and to the provision of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published in print format 2003

ISBN-13 978-0-511-33696-6 eBook (EBL)

ISBN-10 0-511-33696-9 eBook (EBL)

ISBN-13 978-0-521-45506-0 hardback

ISBN-10 0-521-45506-5 hardback

Cambridge University Press has no responsibility for the persistence or accuracy of urls for external or third-party internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

To my children
Rachel, Catherine, and Annalie
who have not known a time when I was
not writing “the book”!

Contents

Preface	xiii
Background: what you need to know before you start	xxiii
1 Gravity on Earth: the inescapable force	1
• Galileo: the beginnings of the science of gravity • The acceleration of gravity is uniform • Trajectories of cannonballs • Galileo: the first relativist	
2 And then came Newton: gravity takes center stage	9
• The second law: weight and mass • The third law, and its loophole • Preview: Newton's gravity • Action at a distance • The new equivalence principle • The gravitational redshift of light • Gravity slows time • Summing up	
3 Satellites: what goes up doesn't always come down	19
• Taking motion apart • Acceleration, and how to change your weight • Getting into orbit	
4 The Solar System: a triumph for Newtonian gravity	25
• How to invent Newton's law for the acceleration of gravity • The orbits of the planets described by Newton's law of gravity • What is the value of G ? • Kepler's laws • The Sun has a little orbit of its own • Geostationary satellites • The gravitational attraction of spherical objects • Playing with the orbit program • Black holes before 1800 • Light is deflected by the Sun's gravity	
5 Tides and tidal forces: the real signature of gravity	39
• Tidal forces in free fall • Ocean tides • Tides from the Sun • Spring and neap tides • What the tidal forces do to the oceans, the Earth, and the Moon • Tides elsewhere in astronomy • Jupiter gives Mercury's story another twist • Triumph of Newtonian gravity: the prediction of Neptune • Tiny flaw of Newtonian gravity: Mercury's perihelion motion	
6 Interplanetary travel: the cosmic roller-coaster	51
• Getting away from the Earth • Plain old momentum, and how rockets use it • Energy, and how planets never lose it • Getting to another planet • The principle of the slingshot • Using Jupiter to reach the outer planets • Slinging towards the Sun • Force and energy: how to change the energy of a body • Time and energy	

7 Atmospheres: keeping planets covered	65
<ul style="list-style-type: none"> • In the beginning ... • ... was the greenhouse ... • ... and then came Darwin • The ones that get away • The Earth's atmosphere • Pressure beats gravity: Archimedes buoys up balloons • Pressure beats gravity again: Bernoulli lifts airplanes • Helium balloons and the equivalence principle • Absolute zero: the coldest temperature of all • Why there is a coldest temperature: the random nature of heat • The ideal gas • An atmosphere at constant temperature • The Earth's atmosphere • The atmospheres of other planets • Quantum theory and absolute zero 	
8 Gravity in the Sun: keeping the heat on	85
<ul style="list-style-type: none"> • Sunburn shows that light comes in packets, called photons • A gas made of photons • Einstein in 1905 • Gravity keeps the Sun round • The Sun is one big atmosphere • The Standard Model of the Sun • The structure of the Sun • How photons randomly 'walk' through the Sun • Rotation keeps the Sun going around • Solar seismology: the ringing Sun 	
9 Reaching for the stars: the emptiness of outer space	103
<ul style="list-style-type: none"> • Leaping out of the Solar System • How far away are the stars? • How bright are stars? • Astronomers' units for brightness • Standard candles: using brightness to measure distance 	
10 The colors of stars: why they are black (bodies)	109
<ul style="list-style-type: none"> • The colors of stars • Why stars are black bodies • The color of a black body • Relation between color and temperature: greenhouses again • Spectral lines: the fingerprint of a star • How big stars are: color and distance tell us the size • But why are stars as hot as they are, and no hotter? • Looking ahead 	
11 Stars at work: factories for the Universe	121
<ul style="list-style-type: none"> • Star light, star bright ... • ... first star I see tonight • Cooking up the elements • The solar neutrino problem • Life came from the stars, but would you have bet on it? 	
12 Birth to death: the life cycle of the stars	135
<ul style="list-style-type: none"> • Starbirth • The gravitational thermostat • The main sequence • Giants • Degenerate stars: what happens when the nuclear fire goes out • The Chandrasekhar mass: white dwarfs can't get too heavy • Neutron stars • Fire or ice: supernova or white dwarf • Death by disintegration • What is left behind: cinders and seeds 	
13 Binary stars: tidal forces on a huge scale	153
<ul style="list-style-type: none"> • Looking at binaries • The orbit of a binary • Planetary perturbations • Tidal forces in binary systems • Accretion disks in binaries • Compact-object binaries • Fun with the three-body problem 	

14 Galaxies: atoms in the Universe	163
• Globular clusters: minigalaxies within galaxies • Describing galaxies • Galaxies are speeding apart • Measuring the Universe: the distances between galaxies	
• Most of the Universe is missing! • Gangs of galaxies • The missing mass • Radio galaxies: the monster is a giant black hole • Quasars: feeding the monster	
• Galaxy formation: how did it all start? Did it all start?	
15 Physics at speed: Einstein stands on Galileo's shoulders	179
• Fast motion means relativity • Relativity is special • The Michelson-Morley experiment: light presents a puzzle • Michelson's interferometer: the relativity instrument • Special relativity: general consequences • The extra inertia of pressure • Conclusions	
16 Relating to Einstein: logic and experiment in relativity	195
• Nothing can travel faster than light • Light cannot be made to stand still • Clocks run slower when they move • The length of an object contracts along its motion • Loss of simultaneity • The mass of an object increases with its speed	
• Energy is equivalent to mass • Photons have zero rest-mass • Consistency of relativity: the twin paradox saves the world • Relativity and the real world	
17 Spacetime geometry: finding out what is <i>not</i> relative	211
• Gravity in general relativity is ... • ...geometry • Spacetime: time and space are inseparable • Relativity of time in the spacetime diagram • Time dethroned ... • ...and the metric reigns supreme! • The geometry of relativity • Proper measures of time and distance • Equivalence principle: the road to curvature ...	
• ...is a geodesic • The equivalence principle: spacetime is smooth	
18 Einstein's gravity: Einstein climbs onto Newton's shoulders	225
• Driving from Atlanta to Alaska, or from Cape Town to Cairo • Dimpled and wiggly: describing any surface • Newtonian gravity as the curvature of time • Do the planets follow the geodesics of this time-curvature? • How to define the conserved energy of a particle • The deflection of light: space has to be curved, too • Space curvature is a critical test of general relativity • How Einstein knew he was right: Mercury's orbital precession • Weak gravity, strong gravity	
19 Einstein's recipe: fashioning the geometry of gravity	239
• Einstein's kitchen: the ingredients • Einstein's kitchen: the active gravitational mass comes first • Einstein's kitchen: the recipe for curving time • Einstein's kitchen: the recipe for curving space • Einstein's kitchen: the recipe for gravitomagnetism • The geometry of gravitomagnetism • Gyroscopes, Lense, Thirring, and Mach • The cosmological constant: making use of negative pressure • The big picture: all the field equations • The search for simplicity • General relativity • Looking ahead	

20 Neutron stars: laboratories of strong gravity	261
<ul style="list-style-type: none"> • Nuclear pudding: the density of a neutron star • It takes a whole star to do the work of 100 neutrons • What would a neutron star look like? • Where should astronomers look for neutron stars? • Pulsars: neutron stars that advertise themselves • The mystery of the way pulsars emit radiation • The rotation rate of pulsars and how it changes • Puzzles about the rotation of pulsars • Pulsars in binary systems • X-ray binary neutron stars • Gamma-ray bursts: deaths of neutron stars? • The relativistic structure of a neutron star • The relation of mass to radius for neutron stars • Neutron stars as physics labs 	
21 Black holes: gravity's one-way street	285
<ul style="list-style-type: none"> • The first black hole • What black holes can do – to photons • The gravitational redshift • Danger: horizon! • Getting away from it all • Singularities, naked or otherwise • What black holes can do ... to orbits • Making a black hole: the bigger, the easier • Inside the black hole • Disturbed black holes • Limits on the possible • The uniqueness of the black hole • Spinning black holes drag everything with them • The naked truth about fast black holes • Mining the energy reservoir of a spinning black hole • Accretion onto black holes • The signature of the supermassive black hole in MCG-6-30-15 • Wormholes: space and time tubes • Hawking radiation: black holes are truly black bodies • Black hole entropy: a link to nineteenth century physics • Black hole entropy: a link to twenty-first century physics 	
22 Gravitational waves: gravity speaks	309
<ul style="list-style-type: none"> • Gravitational waves are inevitable • Transverse waves of tidal acceleration • How gravitational waves act on matter • Early confusion: are gravitational waves real? • How gravitational waves are created • Strength of gravitational waves • Gravitational waves carry energy, lots of energy • The Binary Pulsar: a Nobel-Prize laboratory • Gravitational waves from binary systems • Listening to black holes • Gravitational collapse and pulsars • Gravitational waves from the Big Bang: the Big Prize • Catching the waves • Michelson returns: the relativity instrument searches for waves • LISA: catching gravitational waves in space 	
23 Gravitational lenses: bringing the Universe into focus	331
<ul style="list-style-type: none"> • Pretty obvious, really, ... • ... but not always easy • How a gravitational lens works • Why images get brighter • Making multiple images: getting caustic about light • The Einstein ring • MACHOS grab the light • The third image: the ghost in a mirror • Lensing shows us the true size of quasars • Weak lensing reveals strong gravity 	
24 Cosmology: the study of <i>everything</i>	345
<ul style="list-style-type: none"> • What is “everything”? • Copernican principle: “everything” is the same “everywhere” • The Hubble expansion and the Big Bang • The accelerating Universe • Was there a Big Bang? • Looking back nearly to the beginning • Cosmic microwave background: echo of the Big Bang • The rest frame of the Universe • Big Crunch or Big Freeze: what happens next? • Cosmology according to Newton • Cosmology according to Einstein • Evolving the Universe • The cosmological scale-factor • What is the cosmological expansion: does space itself expand? • The age of the Universe 	

25 The Big Bang: the seed from which we grew	367
• Physical cosmology: everything but the first nanosecond • The expansion of the quark soup and its radiation • The laws of physics prefer matter over anti-matter • The Universe becomes ordinary • Making helium: first steps toward life • Does it correspond to reality? • Three and only three neutrinos: a triumph for Big Bang physics • From nuclei to atoms: the Universe goes transparent • The evolution of structure • Ghosts of the dark matter • What is the dark matter?	
26 Einstein's Universe: the geometry of cosmology	383
• Cosmology could be complicated . . . • . . . but in fact it is simple (fortunately!) • Gravity is geometry: what is the geometry of the Universe? • Friedmann's model universes • What the Universe looks like	
27 Ask the Universe: cosmic questions at the frontiers of gravity	391
• The puzzle of the slightly lumpy Universe • Einstein's "big blunder" • The cosmological constant in particle physics • Inflation: a concept waiting for a theory • Inflation power: the active vacuum • Inflating the Universe • Inflation put to the test • Is inflation still going on? • Is Einstein's law of gravity simply wrong? • Cosmic defects • Cosmic rays • Quantum gravity: the end of general relativity • A Universe for life: the Anthropic Principle • Causality in quantum gravity: we are all quantized • The quantization of time? • Time for the twenty-first century	
Appendix: values of useful constants	419
Glossary	421
Index	443

Preface

From the author to the reader

Why this book is about gravity

During the 35 years that I have done research in gravitation, I have watched with amazement and delight as my colleagues in astronomy have, step-by-step, opened up almost the entire Universe to our view. And what a view! There are punctures in space called black holes that capture gas and stars with a relentless and unbreakable grip; there are 10 km balls called neutron stars that are immense overgrown atomic nuclei with more mass than our Sun, that spin about their axes hundreds of times per second while emitting intense beams of radiation; there are bursts of gamma-rays from the most remote regions of the Universe that are so intense that they outshine the rest of the Universe for a short time; and most strikingly of all there was the beginning of time itself in an explosion of pure energy, driven by a force we do not understand, in which matter as we know it did not exist, in which even the laws of Nature themselves were mutable.

This astonishing Universe has captured the imagination of many people, among them many scientists. Physicists trained in a number of disciplines have applied themselves to explaining these and many more less spectacular but equally important phenomena, such as: how the chemical elements were made; where stars come from and how they evolve and die; how the vast systems of stars called galaxies formed and why they have grouped themselves into clumps and long chains; why a Universe filled with bright stars seems to contain even more matter that cannot form stars – and so remains dark.

From all this scientific activity has come a great deal of understanding. We know not just *what* happens, but in many cases *how* and *why*. Physicists, astrophysicists, and astronomers have been able to put together a coherent story of how our Universe began and of how its immense variety evolved.

The central theme of the story of the Universe turns out to be *gravity*.

Gravity is the one force of Nature that operates everywhere; it controls the effects of all the other forces wherever they act; it regulates countless natural clocks, from the orbits of planets to the lifetimes of stars. Gravity rules the most violent places in the Universe – quasars, pulsars, gamma-ray bursters, supernovae – and the most quiet – black holes, molecular clouds, the cosmic microwave background radiation. Today gravity binds stars and galaxies and clusters of galaxies together, but much earlier it pushed the Universe violently apart. Gravity explains the uniformity of the Universe on very large scales and its incredible variety on small scales. Gravity even laid the path toward the evolution of life itself. If we understand how gravity works, then we begin to understand the Universe.

Rich as our understanding of the workings of gravity in the Universe has become, it is far from complete. The gaps are not just hidden regions, phenomena yet to be discovered, although when such discoveries occur they are sure to bring more

►The link between gravity and the wonders of astronomy goes right back to Galileo, who founded the science of gravity. Using a telescope for the first time, Galileo became the first person to understand that the Milky Way is composed of stars, that Venus shines by light reflected from the Sun, that the Sun is plagued by spots, that Jupiter holds its own satellites in orbit around itself in imitation of the Solar System. Our amazement at astronomers' discoveries today helps us to appreciate what Galileo's contemporaries must have felt at his.

▷The word “revolution” gets used so much these days in discussions of progress in the sciences, that I hesitate to use it here. But I know no better word.

▷Our tour will be thought-provoking, sometimes demanding, even laborious. But we will not leave you, the reader, behind. If you start with high-school mathematics skills – see the section “How this book uses mathematics” beginning on the next page – then you will be able to follow the discussion all the way. And if you put in the effort to study and run the computer programs that allow you to study areas where simple mathematics does not suffice, then you can reach real expertise in some areas. Our goal is ambitious, but I hope you find it worth the effort!

amazement and delight. The most exciting gaps are those in our understanding of the laws of Nature.

The enormous advances in astronomy that I have witnessed in my working life have brought us to the threshold of a profound revolution in our understanding of gravity itself. Many physicists today are working to unify gravity with the other forces of Nature, which will lead to what is called a quantum theory of gravity. There are aspects of the Universe that will not be explained without this new theory, and there are clues to the new theory in many of the currently unexplained puzzles of the Universe.

This book is about gravity at the threshold of this revolution. We will take a tour of the Universe from the ground up. We will start at the surface of the Earth and move outwards through the Solar System, the Galaxy, and beyond to a scale where our Galaxy is the merest atom in the corpus of the Universe.

We will learn about gravity and the other laws that govern the Universe, first as understood by Newton and his successors, then as understood by Einstein and modern physicists. We will use these laws to see how the parts of the Universe work, how they relate to one another, and how they may have come to be. By the end of our tour we will see the Universe and its physical laws, not merely as a collection of fascinating but separate phenomena, but rather as a unity.

Our goal is not just to wonder and marvel at our Universe, nor simply to admire the cleverness of the scientists who have made the Universe at least partially understandable. Instead, *our goal is to understand how the Universe works, to begin to think about the Universe in the same way that these scientists themselves do.*

How gravity evolves

Gravity, the oldest force known to mankind, is in many ways also the youngest. It is understood well enough to explain stars, black holes and the Big Bang, and yet in some ways it is not understood at all. Explaining gravity required the two greatest scientific minds of modern history, Isaac Newton and Albert Einstein; and now hundreds of the brightest theoretical physicists are working to invent it once again. Each time gravity has been re-invented, it has sparked a revolution. Newton’s theory of gravity stimulated huge advances in mathematics and astronomy; indeed, it was the beginning of modern theoretical physics. Einstein’s theory of gravity, which he called general relativity, opened up completely unexpected phenomena to investigation: black holes, gravitational waves, the Big Bang. When, sometime in the future, gravity changes into quantum gravity, possibly becoming just one of many faces of a unified theory of all the physical forces, the ensuing revolution may be even more far-reaching.

Each of these revolutions has built on the previous one, without undermining it. Newton’s gravity is just as important today for explaining the motions of the planets as in Newton’s time. It is used to predict the trajectories of spacecraft and to understand the structure of galaxies. Yet general relativity underpins all of this, because Newton’s gravity is only an approximation to the real thing. We need only Newton to help us understand how a star is born and evolves; but when the star’s evolution leads to gravitational collapse and a supernova, then we have to ask Einstein’s help to understand the neutron star or black hole that is left behind. When we have a theory of quantum gravity, it won’t stop us from using general relativity to explain how the Universe expanded after the Big Bang; but if we want to know

where the Big Bang came from, and why (or whether) time itself started just then, we will need to ask the quantum theory.

There is a deeper reason for this continuity from one revolution to the next. As an example, consider the fact that two of the fundamental ideas in Einstein's general relativity, called the principle of relativity and the principle of equivalence, originated with Galileo. Einstein's revolution brought a complete change in the mathematical form of the theory, added new ideas, and opened up new phenomena to investigation. But there was a profound continuity in physical ideas, and these were as important to Einstein as the mathematical form of the theory. The coming quantum revolution will surely likewise be grounded firmly in concepts that physicists today use to understand gravity. These physical ideas are the subject of this book.

Of course, this book deals mainly with what we already know about the role that gravity plays in the Universe, which is the result of the first two revolutions. You, the reader, will learn what Newton's gravity is, and how it regulates planets, stars, and galaxies. You will learn what relativity means, and how general relativity leads to black holes and the Big Bang. But I want you also to be able to follow the continuity of ideas, to see for example how Newtonian gravity prepared the way for relativity. In the earliest chapters we will see that Newtonian gravity already contained half of relativity, that it contained the equivalence principle that guided Einstein to general relativity, even that it foresaw the existence of black holes and gravitational lenses. In the same way, general relativity contains seeds that will blossom only when the third revolution arrives, such as why the theory allows the cosmological constant. I will try to point out some of these seeds as we go along, usually by asking questions that general relativity or modern astronomy suggests but does not answer. The final chapter is devoted entirely to such questions.

Why this book is about more than gravity

Because gravity is the dominant force anywhere outside of the surface of the Earth, this book covers a lot of astronomy. But instead of just touring randomly around the Universe, we have a theme: gravity as the engine that makes things happen everywhere. This theme unifies and simplifies the study of astronomy. If we understand gravity on the Earth, then it is easier to understand it in the Solar System. If we understand it in the Solar System, then we have an easier time grasping how it acts in stars and black holes. And so it goes, right up to the largest scales, to the Universe as a whole.

Because gravity usually acts in concert with other forces of physics, studying gravity this way also gives us the opportunity to investigate much of the rest of physics along the way. For example, quantum theory and gas dynamics play important roles in stars, and so we study them in their own right where we need them. Even if you have studied these subjects before, you may be surprised at some of the connections to other parts of physics that you will discover by looking at them in the context of explaining a star.

How this book uses mathematics

You may already have guessed that this book is not a "gee-whizz" tour of the Universe: this is a book for people who are not afraid to think, who want to understand what gravity is, who want to go beyond the superficial level of understanding that many popular books settle for. But this is also not an advanced textbook. We shall steer a careful middle course between the over-simplification of some popular treatments and the dense complexity of many advanced mathematical texts.

This book has equations, but the equations use algebra and (a little) trigonom-

▷ Chapter 1
▷ Chapter 2
▷ Chapter 4
▷ Chapter 27
▷ Chapter 27

▷ Chapters 1–3
▷ Chapters 4–8
▷ Chapters 9–13, Chapter 21
▷ Chapters 24–27

▷ Beginning in Chapter 7

etry, not advanced university mathematics. What is required in place of advanced mathematics is thought: readers are asked to reason carefully, to follow the links between subjects. You will find that you can climb the ladder from gravity on the Earth to gravity (and even anti-gravity) in the Universe if you go one step at a time, making sure you place each foot securely and carefully on the rungs as you climb. In return for putting in the thought that this book asks, you can get much further than you might have expected in understanding gravity and its manifestations in astronomy. School students and university undergraduates will find that this book offers them an early avenue into subjects that are usually regarded as much too advanced for them.

There is no calculus in this book, despite the fact that calculus is the workhorse mathematical tool of physics. Wherever possible I have tried to present a physical argument as a substitute for the mathematical one that physicists are used to. This has the great advantage that it makes connections between different parts of physics clearer and the logical reasoning more direct. It has the disadvantage, of course, that it is not always possible to do this: there are places where using more advanced mathematics really is necessary for a pen-and-paper treatment. In such cases I have often turned instead to a computer program. These programs are not “black boxes”: their construction is discussed in detail. See the next section for a discussion of why they are good substitutes for advanced mathematics.

Sometimes I have had to resort to that awful phrase “it can be shown with more advanced mathematics” or something like it. I have avoided this whenever I could, but there are times when it seemed to me that any argument I could give for a particular result would be over-simplified, it would hide or corrupt the truth. It is best in such situations to be honest and accept that our mathematical tools at this level are not always sufficient. Our aim is not to cut corners, but always to remain true to the physics.

In fact, it is possible to read this book while avoiding most of the equations, if you want to. All the extended algebraic calculations are placed in special boxes, called investigations. These are set aside on a light-gray background. Skipping these boxes might be a good strategy if you are short of time, or on your first reading. If you skip them you will just have to take on faith some of their results, which are then used in the main text. Many of the investigations contain exercises, which offer you a chance to test your understanding. I believe strongly that doing exercises is the most effective way to get comfortable with an important result. If you are using this book as a textbook for a course, then I hope your teacher will expect you to do the exercises!

How to go beyond this book by using computers

For those of you who have access to computers and want to use them, I have provided a way to reach the results of some very advanced mathematics by using computer programs that only require the mathematical level of the rest of this book. This is your best way to get to some of the results that algebra alone cannot reach. The programs can be downloaded from the website (see the next section) and used right away – just run them and look at the results. You can then change some of the numbers they work with, for example to compute the orbit of Jupiter instead of Mercury, without looking inside the program. But the way to get the most out of them is to study the investigations in which they are described, look inside the programs, and even experiment with changing the code.

As an example of the power of computer programs, consider the motion of a planet around the Sun. Newton’s law of gravity giving the forces that govern the motion of the planet is not hard to write down or to understand using pure algebra.

▷ See Figure 19.1 on page 242.

▷ Solutions to the exercises can be found on the book’s website. See the next page.

Using it – solving it – to find the planet’s orbit is not so easy with pen and paper. The usual way that university physics students learn how to show, for example, that the orbit is an ellipse is by using some rather sophisticated calculus. They may have to wait until their second or third year to get to this important result. In Chapter 4 we achieve the same thing by writing a simple computer program that moves the particle along in its orbit step-by-step. The law of gravity translates directly into a prescription for algebraic calculations that the computer can readily do. The orbit that comes out is clearly an ellipse. The calculation can be done as accurately as one wishes by simply telling the computer to take smaller steps. Repetitious computations like this are what computers do best.

Even better, a computer program can be modified and used again in situations that would require even more sophisticated calculus to make progress. We modify the orbit program slightly in Chapter 13 to explore what happens when three stars interact with one another, a situation that often results in one of them being expelled from the system at high speed. We modify the orbit program yet again in Chapter 21 to show that orbits of bodies around black holes are not ellipses, but rather make rosette patterns. And finally in Chapter 24 we modify it another time to calculate the expansion of the Universe itself.

The programs are available from the website of this book (see below). They are written in the free and widely available language Java®. Since the popularity of this language is steadily increasing, you may find that following the computer programs here will help you to learn a language that will be useful to you later. In any case, the Java language is not very different from the most popular language of all, C. If you have never programmed before, the package you can download from the website provides an easy way of starting.

Using the glossary and website

There are two aids for you to get more information than is in the text: the glossary, which begins on page 421, and the website. The glossary contains definitions of many of the terms used in this book. Some of the terms in the glossary are not defined in the book because I have assumed that most readers will know what they mean, or because they are not central to the subject matter. Other terms in the glossary are important words that are explained somewhere in the book but which are placed in the glossary so that you can conveniently look them up when you encounter them again. All terms in the glossary are printed in **boldface** type when they are first used in the book. I don’t use boldface for any other reason, so whenever you see it you will know that it marks an entry in the glossary.

The website for this book is

<http://www.gravityfromthegroundup.org>

It contains

- the Java programs for you to download;
- a free version of the Triana® software environment for running the programs and displaying their results graphically;
- solutions of all the exercises;
- links to allow you to download and install Java and other programs needed for your computer;
- additional illustrations for some of the chapters;
- a way of submitting comments, misprints you have found, or suggestions that could be incorporated in future editions; and

►This so-called three-body problem is a classic problem that cannot be solved by analytic calculus alone. All scientists who study it use computer programs.

►Terms printed in **boldface** type in the text are contained in the glossary.

- links to useful websites where you can follow up some of the material covered in the book.

Visit the website: it is a valuable addition to this book, and it is completely free.



From the author to his colleagues

Teaching gravitation

Although this book is aimed at beginners, many readers may be my colleagues, professional physicists who may be using the book in a course they are teaching, supervising a student who is using the book for a self-study program, or just looking for a different point of view on the subject. For such readers, this section enlarges on the pedagogical side of my approach to this subject.

The aim of this book is to introduce gravitation theory as a unified subject, and especially to show the key role that gravitation plays in the phenomena of the Universe. An associated pedagogical goal is to develop the reader's ability to think physically by using physical reasoning rather than advanced mathematics to move through the subject. The restriction to elementary mathematics presents real challenges in presentation, but it allows one to treat the entire theory in a unified way, from Newton to Einstein and beyond ... from the ground up, in other words.

Mathematics is not just a powerful tool in physics, it is the *reference language* of science: it is the language in which the fundamental theories are written, the medium which is used to deduce the predictions from a theory. But physicists generally supplement mathematical deduction with physical reasoning. Indeed, physicists who can do this reliably are widely admired for their great "physical intuition". When physicists are inventing new theories, searching for new physics, or trying to explain phenomena not previously encountered, physical reasoning often leads and mathematical reasoning follows: new ideas suggested first by physical arguments are then put into mathematical form and tested for consistency and suitability. Yet when physicists teach known physics to newcomers, the balance more often falls heavily toward mathematical reasoning, the reference form of the theory. Students need of course to master the mathematical form of a theory in order to be able to work seriously with it or to go beyond it, and physics teaching generally focuses on that requirement.

But I believe that this focus is often too narrow. It is important to remind ourselves that there is usually a line of physical reasoning that moves along parallel to the mathematical. Ideally, each way of thinking supports the other. But for students with unsophisticated mathematical tools, it should be possible to make significant progress using mainly physical reasoning. After all, if the principal theories of physics were invented by using physical arguments as guides, then it should be possible to teach important things about those theories in the same way.

Putting mathematical presentation first has another undesirable pedagogical side-effect: it is customary to teach some physical theories in discontinuous segments in order to allow students time to learn more sophisticated mathematics in between. Nowhere is this more arresting than for gravitation theory. Newton's law of gravity is presented in high school, but even using it to find the simple elliptical orbit of a planet must wait until the student masters integral calculus, in the first or second year of a university course. Because of its use of tensors and differential geometry, general relativity has to wait until the final undergraduate year at the earliest; most physicists encounter it first as graduate students, if at all.

Yet there are very good reasons for teaching gravitation theory as a unified subject. The continuity of physical ideas and phenomena is strong. Consider the following sampling, which is by no means exhaustive.

- The equivalence principle and the principle of relativity – so important to Einstein – originated with Galileo.
- Most physicists find it remarkable that black holes and the gravitational deflection of light were discussed by scientists more than a century before Einstein (see Chapter 2). Yet surely this simply means that the links between Newtonian and Einsteinian gravity go deeper than most of us assume.
- There are more similarities than differences between Newtonian and relativistic stars. Even the gravitational effects caused by the spins of stars and black holes have their roots in Newtonian gravity (Chapter 19).
- If we want to trace the histories of the objects in Newton’s universe, such as planets and people, all the way back to their ultimate roots, we inevitably encounter the hot, slightly lumpy plasma that we call the Big Bang. The inevitability of the Big Bang has as much to do with the argument of the eighteenth-century physicist Olbers, who understood how strange it is that the sky is dark at night, as it has with Einstein.
- And even Einstein’s theory of general relativity itself, for all its mathematical complexity, is arguably as close in physical content to Newton’s as it was possible for Einstein to make it while still respecting special relativity.

To teach the broad sweep of gravity as a unified whole to an audience that normally only gets taught about circular planetary orbits, I have followed the pedagogical philosophy outlined earlier: using the minimum level of mathematical sophistication, I have tried to progress through gravitation theory as much as possible by using physical arguments. This started out, quite frankly, as an experiment, a challenge to myself, and I have learned much from it, especially about the connections between and continuity of ideas in this subject. For example, it is satisfying that it is natural to introduce both the principles of relativity and of equivalence in Chapter 1, followed immediately by the gravitational effect on time in Chapter 2, without ever leaving the vicinity of the Earth, before even considering Newton’s law of gravitation. When these principles turn up again in special and general relativity, they are old friends. When I explain in Chapter 19 that gravity in Einstein’s picture is found mainly in the curvature of time, it is not hard to justify this from the discussion in Chapter 2. It is equally satisfying to calculate the Newtonian gravitational force exerted by a spherical body (Chapter 4) – using one of the computer programs to do the integral calculus – and then to find that one needs nothing more than this to calculate the evolution of a homogeneous and isotropic cosmological model (Chapter 24). It is fascinating to calculate the fundamental normal mode frequency of the Sun (Chapter 8) in Newtonian gravity and then to find that the same formula comes within a factor of two of the right answer for the pulsations of a disturbed black hole (Chapter 21). It is equally fascinating to discover that one can derive the Lense–Thirring effect quantitatively from Newtonian gravity and special relativity only if one uses the Einstein form of the active gravitational mass as the source of gravity (Chapter 19), and thereby to establish deep links between Newtonian gravity, the spinning black holes (Chapter 21), and the inflationary universe (Chapter 24). The list could be much longer.

This approach to teaching gravitation will surely not appeal to everyone, but I hope that especially my scientific colleagues in relativity will find it amusing to see how many threads continue from Newtonian gravity to relativity, how many apparently abstract and mathematical properties of relativistic gravity have clear and simple physical derivations, and how much easier it is to introduce general relativity if Newtonian gravity has been taught in a way that emphasizes the ideas that continue into relativity.

Guiding students through this book

The book can be used by teachers for guided self-study, as a textbook for a course on physics or astronomy for non-scientists, or as a main or supplementary text in conventional university physics and astronomy courses.

Courses for non-scientists need to excite and challenge students without overwhelming them with mathematics. With its emphasis on developing physical intuition, this book aims directly at what is probably the most important goal of such a course: students should learn what it means to think like a physicist. The exercises can play an important role. Depending on the length of such a course and the background of the students in it, the teacher may want to be selective in what material to focus on. I would welcome feedback (via the website) from anyone teaching such a course.

When using this book with physics or astronomy undergraduates, the obvious problem is that the book has a “vertical” integration: it covers material that is usually treated in different courses in different years. Indeed, that is why I have written it. It can be helpful for beginners to expose them to some of these advanced ideas and to let them explore them with the aid of the computer programs before they reach the mathematical level needed to treat them in the conventional way, later in their education. Again I would welcome feedback via the website from lecturers who use this book in such courses, either as the main or as a supplementary text.

The computer programs deserve special attention from the teacher. They fill the gaps between algebra and calculus for beginners, while for students who continue to study physics they are good preparation for later analytical attacks on problems.

►Besides the four equations-of-motion computer problems mentioned earlier, the book applies computers to a variety of other problems. Students can prove that the gravitational field outside a spherical body is the same as if all its mass were concentrated at its center, by adding up the forces from small elements of the body. They can make a computer model of the Earth’s atmosphere, and later use the same program to model the Sun and a neutron star. In each case the problems are formulated from the start in terms of small differences rather than derivatives. And there are no compromises: we don’t have to over-simplify these problems in order to put them on the computer.

Let me give an example. Consider the computer program for finding the motion of a planet around the Sun, to which I referred earlier. The mathematical way that undergraduate physics students learn that its orbit is an ellipse is by writing Newton’s law of motion as a differential equation and solving it using fairly sophisticated calculus. They often have to wait a year or two in their undergraduate course before they have the skill to do this. The solution can instead be found using a computer if we replace the differential equations with finite difference equations. By formulating Newton’s law from the start in terms of finite differences – the change in velocity in a small but finite time-interval is approximately the acceleration at the beginning of the time-interval times the time-interval – we have an immediate entry into the computer simulation. The formulation is obvious to students, and just as obvious is the idea that if one makes the time-step smaller and smaller then the computer solution becomes a better and better approximation to the real thing. This is calculus in practice, and if students meet calculus later in their mathematics education, then they know they have already been doing it on the computer.

Computers are already used in this way in many introductory physics courses. Some of the best use spreadsheets, because they are widely available, they contain all the required mathematical operations, and they can display results as graphs. For this book, however, I have chosen instead to use the programming language Java®. It is also widely available, free, and mathematically complete. The Triana® environment that can be downloaded for free from the website provides the ability to run pro-

grams as black boxes and get graphical output. I prefer the fact that Java programs are closer in structure to those that students may write later in their careers (in Fortran, C, or even Java). But lecturers who already use spreadsheets with their students should have little trouble transferring the programs to that format.



Acknowledgements

This book has taken many years to write, and in that time I have learned much from many colleagues in astronomy, physics, and relativity. Some have contributed directly to this book with constructive criticism, creative input, or tracking down resources and historical material; others have simply taught me things I did not know. I would especially like to acknowledge my indebtedness to Robert Beig, Werner Benger, Jiří Bičák, Curt Cutler, Thibault Damour, Karsten Danzmann, Mike Edmunds, Jürgen Ehlers, Jim Hartle, Günther Hasinger, Jim Hough, Klaus Fricke, Matthew Griffiths, Geraint Lewis, Elke Müller, Charlie Misner, Jürgen Renn, Rachel Schutz, Ed Seidel, Kip Thorne, Joachim Wambsganss, Ant Whitworth, Chandra Wickramasinghe, and Cliff Will. None of them, of course, bears responsibility for any errors that remain in the book. My employers, the Max Planck Society and Cardiff University, have kindly made it possible for me on a number of occasions to get away from my normal duties and write. My editors at Cambridge University Press – Simon Mitton, Rufus Neal, Simon Capelin, Tamsin van Essen – deserve special thanks for their patience and encouragement while waiting for a manuscript that must have seemed like it might never arrive, and which kept growing well beyond its original planned length. Fiona Chapman of Cambridge University Press helped enormously with the presentation. And last but by no means least, I want to thank my family, especially Siân, for putting up with my hiding away to write on countless evenings, weekends, and holidays, and for their unwavering belief that the final result would be worth it.



Ready to start

This preface is long enough. Beginners, or rusty old-timers, should check the review of background material that follows next; others should jump straight to Chapter 1. I wish the reader a satisfying and enlightening journey through the universe of gravity, from the ground up.

Bernard Schutz
Golm, Germany
16 February 2003

Background: what you need to know before you start

As explained in the preface, I have used high-school mathematics to present some of the material in this book. If you want to know what that means, if you want to learn whether you have the background necessary to do the mathematics, then scan through this introductory material. But remember, it is not necessary to follow all the derivations, particularly the ones in the boxes, if you just want to learn what the main ideas in modern gravity and astronomy are. So if you find your mathematics too old or rusty, then see how you get along without it.

High-school mathematics

The mathematics used is basic numeracy, algebra, and a tiny bit of trigonometry (which you can skip).

It is essential to understand scientific notation for numbers, that is how to write numbers in the form 3.2×10^6 and know what the factor 10^6 means. Scientists use this notation all the time, because otherwise they would be writing out long confusing strings of zeros. The number 3.2×10^6 means 3 200 000, obtained by moving the decimal point in 3.2 six places to the right. Similarly, the number 5.9×10^{-3} is 0.0059, obtained by moving the decimal point three places to the left.

Scientific notation also allows scientists to hint at the accuracy with which they know a number, or at least intend to use it. So if a scientist measures a brief lapse of time to the nearest thousandth of a second and gets 0.021 seconds, then he can write it as 2.1×10^{-2} s. If the measurement accuracy were greater, say it came out to be 0.0210 accurate to one ten-thousandth of a second, then the scientist could write 2.10×10^{-2} s. The number of figures quoted before the power of ten is called the number of *significant figures* in the expression. Generally, in working out calculations, you should aim to keep only as many significant figures in your answer as there were in the least accurately known number you used in your calculation. There is no point using π to ten figures to compute the area of a circle if you have measured its radius to only 10% accuracy.

Problems in physics often involve numbers that have *units*, like seconds or meters. We will use SI units in this book, and the values of important physical numbers in these units are given in the Appendix. Some readers may be more comfortable with American or British Imperial units, like feet, miles, pounds. But all fundamental scientific work is done these days in SI units, so I will assume that you know the conversions.

I also assume you know the basics of algebra. For example, you know how to manipulate and solve simple equations, for example:

$$a + b = c \Rightarrow a = c - b.$$

I often make remarks in the book that we will “solve for a variable” in an equation, as we have done for a above. This means adding, subtracting, dividing, multiplying, or whatever is needed in order to isolate the variable on the left-hand side of the

equation. But we will not need to use advanced solution techniques, like solving quadratic or cubic equations.

When dealing with numbers that have units in algebraic expressions, treat the units as if they were variables. That is, if you square a length of 10 m to get the area of a square, then you square not only the number but also the unit, so that the result is $10^2 \text{ m}^2 = 100 \text{ m}^2$, or 100 square meters. We shall always use negative exponents on units to denote division, where in words we would use “per”: a speed of 15 meters per second is written as 15 m s^{-1} , and an acceleration of 9.8 meters per second per second is written as 9.8 m s^{-2} . Remember as well that, while you can freely multiply numbers that have different units (and get a result whose units are the product of the units of all the factors), you can’t add numbers with different units at all. To add 2 seconds to 3 meters is nonsense. If you find yourself doing this in a calculation, you have made a mistake! Go back and find it.

Angles are always interesting. The everyday unit for angles is the degree, of which there are 360 in a circle. Scientists occasionally use this too, if they have an angle with a convenient size, like 90° . But the standard measure for angles in more advanced mathematics is the radian, and this pops up occasionally in this book as well. The radian measures essentially the fraction of a full circle that the angle represents, except that it expresses this as a fraction of 2π . So a full circle has 360° or 2π radians. One radian is therefore $360/2\pi$ degrees, or about 57.3° . The measure is convenient because it is the ratio of the length of the arc of a circle that the angle intercepts to the radius of the circle.

Being the ratio of two lengths, the radian is a *dimensionless* number. Dimensionless (or pure) numbers have a particular importance in mathematics. There are certain mathematical operations that can be done only with dimensionless numbers, such as evaluating the expression $x + x^2$. We will meet some situations like this in this book, especially when we introduce the exponential function.

Physics

► Terms in **boldface** are defined in the glossary.

This book is about introducing advanced ideas in physics in a simple way, so it helps if the basis for these ideas is already present. I assume that you have had an elementary introduction to physical science, so that we have some common language and ideas. Remember to check the glossary for any terms or words that you are unsure of, especially when you first encounter them written in **boldface** type.

For example, the study of mechanics is the science of how things move under forces. The basis of mechanics is the set of three laws of motion of Newton. We will discuss these, but it helps if you review them beforehand. The first law says that objects move in straight lines if there are no forces acting on them. The second is usually expressed as the equation $F = ma$, which relates the acceleration a of a body to the force applied to it F and its mass m . The third law is the one that causes most grief in first discussions of mechanics: to every force there is an equal and opposite reaction. Since this is discussed in some detail in Chapter 2, don’t worry too much about it before.

It helps if you have looked at Hooke’s law for springs, which states that the force with which a spring pulls back is proportional to the length by which it has been stretched. It also helps if you have encountered the definitions of momentum and, less critically, of angular momentum. But none of these is a show-stopper for this book.

I also assume that the basics of electricity and magnetism are familiar. For example, electric charges come in two types: opposite charges attract each other and similar charges repel. Magnetism similarly has two poles, and is created by moving electric charges.

We shall learn a lot in this book about the details of atoms and quantum theory, so it helps to review some basic facts. All matter is composed of atoms, and each atom has a dense nucleus (made of protons and neutrons) surrounded by orbiting electrons. The electrons are responsible for chemistry: atoms bind to one another to form molecules by the forces that attract the electrons of one atom to the nucleus of another. The number of electrons equals the number of protons in a normal atom, but if electrons are removed it is called an ion. The number of protons determines the kind of element that the atom belongs to, while the number of neutrons can vary. Two nuclei that have the same number of protons but different numbers of neutrons are called different isotopes of the same element. We will use the notation ^{238}U to represent this information: the symbol U tells us we are dealing with uranium, and the prefix 238 represents the total number of protons and neutrons in the nucleus, so it tells us which isotope we have.

Light plays a central role in astronomical observations, so its properties occupy much of this book. By passing light through a prism you can split it into its component colors. This is called its spectrum. Color corresponds to the wavelength or frequency of the light. White light, such as is produced by a light bulb, has a smooth distribution of intensity across the colors of its spectrum. But sometimes the intensity is concentrated at particular colors. Light from a fluorescent light is like this. These “spectral lines” are signatures of the atoms emitting the light. We will go into this in some detail in the book.

Other kinds of radiation are related to light; they are generically known as electromagnetic waves. Radio waves and microwaves are the same thing, only with longer wavelengths. They are usually produced by making electrons accelerate in an antenna; accelerating charges emit electromagnetic radiation. X-rays and gamma-rays have very short wavelengths compared to visible light.

Naturally, it helps if you have some background in gravity. Although we will introduce almost everything we need, it makes things easier if you have seen some things before. For example, the history of gravity is an important issue in the history of ideas: Aristotle’s view that objects fall to the ground because they are seeking their natural place was a big obstacle when Galileo began to formulate laws of motion and gravity mathematically. And Newton’s law of gravity, that the force between two objects was proportional to the product of their masses and inversely proportional to the square of the distance between them, was a revolutionary step whose importance we shall discuss in the book. Finally, everyone knows that Einstein invented relativity and the formula $E = mc^2$ (which you do not need to know the meaning of before you read this book). But he also invented general relativity, which is the theory of gravity that physicists and astronomers use now, replacing Newton’s. The last half of the book is devoted to studying general relativity and its applications in astronomy and fundamental physics.

This book takes you on a tour of the Universe, but again it helps if you have some idea of where we are going. The Earth is one small planet orbiting a modest star called the Sun. The Sun is one of 10^{11} stars that make up the Milky Way galaxy. This is a spiral galaxy, one of perhaps 10^{11} galaxies that are within the reach of our telescopes. The immensity of it all is astonishing.

With modern telescopes astronomers can see extremely far away. Since they use light, which has been traveling at the speed $c = 3 \times 10^8 \text{ m s}^{-1}$, this means they are seeing distant objects as they were at an earlier time. Today astronomers can look back in time most of the way to the beginning of time in the Big Bang. The Big Bang led to the Universe that we see, full of galaxies rushing away from one another. But don’t worry about this: we will deal with it thoroughly in the text.

Computing

The computer programs and exercises are optional, and everything you need to know about them can be downloaded from the website. You don't need to be a computer programming whiz, but you need to have had some experience with computer languages to understand the programs. However, to run the programs and see the results, play with the different possibilities – anyone can do this.

If you are comfortable with these things, then let's not waste any more time thinking about how to learn about gravity. Let's do it.

Gravity on Earth: the inescapable force

Gravity is everywhere. No matter where you go, you can't seem to escape it. Pick up a stone and feel its weight. Then carry it inside a building and feel its weight again: there won't be any difference. Take the stone into a car and speed along at 100 miles per hour on a smooth road: again there won't be any noticeable change in the stone's weight. Take the stone into the gondola of a hot-air balloon that is hovering above the Earth. The balloon may be lighter than air, but the stone weighs just as much as before.

This inescapability of gravity makes it different from all other forces of nature. Try taking a portable radio into a metal enclosure, like a car, and see what happens to its ability to pick up radio stations: it gets seriously worse. Radio waves are one aspect of the *electromagnetic force*, which in other guises gives us static electricity and **magnetic fields**. This force does not penetrate everywhere. It can be excluded from regions if we choose the right material for the walls. Not so for gravity. We could build a room with walls as thick as an Egyptian pyramid and made of any exotic material we choose, and yet the Earth's gravity would be right there inside, as strong as ever. *Gravity acts on everything the same way.*

Every body falls *toward* the ground, regardless of its composition. We know of no substance that accelerates *upwards* because of the Earth's gravity. Again this distinguishes gravity from all the other fundamental forces of Nature. **Electric charges** come in two different signs, the "+" and "-" signs on a battery. A negative **electron** attracts a positive **proton** but repels other electrons.

There is a simple home experiment that will show this. If you have a clothes dryer, find a shirt to which a couple of socks are clinging after they have been dried. Pulling the socks off separates some of the charges of the molecules of the fabric, so that the charges on the sock will attract their opposites on the shirt if they are held near enough. But the socks have the same charge and repel each other when brought together.

The existence of *two* signs of electric charge is responsible for the shape of our everyday world. For example, the balance between attraction and repulsion among the different charges that make up, say, a piece of wood gives it rigidity: try to stretch it and the electrons resist being pulled away from the protons; try to compress it and the electrons resist being squashed up against other electrons. Gravity allows no such fine balances, and we shall see that this means that bodies in which gravity plays a dominant role cannot be rigid. Instead of achieving equilibrium, they have a strong tendency to collapse, sometimes even to **black holes**.

These two facts about gravity, that it is ever-present and always attractive, might make it easy to take it for granted. It seems to be just part of the background, a constant and rather boring feature of our world. But nothing could be further from the truth. Precisely because it penetrates everywhere and cannot be cancelled out, it

In this chapter: the simplest observations about gravity – it is universal and attractive, and it affects all bodies in the same way – have the deepest consequences. Galileo, the first modern physicist, founded the equivalence principle on them; this will guide us throughout the book, including to black holes. Galileo also introduced the principle of relativity, used later by Einstein. We begin here our use of computer programs for solving the equations for moving bodies.

▷Remember, terms in **boldface** are in the glossary.

▷The picture underlying the text on this page is of the famous bell tower at Pisa, where Galileo is said to have demonstrated the key to understanding gravity, that all bodies fall at the same rate. We will discuss this below. Photo by the author.

is the engine of the Universe. All the unexpected and exciting discoveries of modern astronomy – **quasars**, **pulsars**, **neutron stars**, black holes – owe their existence to gravity. It binds together the gases of a **star**, the stars of a **galaxy**, and even galaxies into **galaxy clusters**. It has governed the formation of stars and it regulates the way stars create **chemical elements** of which we are made. On a grand scale, it controls the **expansion of the Universe**. Nearer to home, it holds planets in orbit about the Sun and satellites about the Earth.

The study of gravity, therefore, is in a very real sense the study of practically everything from the surface of the Earth out to the edge of the Universe. But it is even more: it is the study of our own history and evolution right back to the **Big Bang**. Because gravity is everywhere, our study of gravity in this book will take us everywhere, as far away in distance and as far back in time as we have scientific evidence to guide us.

Galileo: the beginnings of the science of gravity

In this section: Galileo laid the foundations for the scientific study of gravity. His demonstration that the speed of fall is independent of the weight of an object was the first statement of the principle of equivalence, which will lead us later to the idea of black holes.

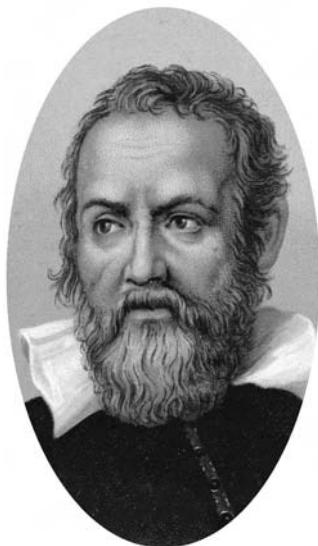


Figure 1.1. Galileo Galilei moved science away from speculation and philosophy and toward its modern form, insisting on the pre-eminence of careful experiment and observation. He also introduced the idea of describing the laws of nature mathematically. Meeting strong religious opposition in his native Italy, his ideas stimulated the growth of science in northern Europe in the decades after his death. Image reproduced courtesy of Mary Evans Picture Library.

We will begin our study of gravity with our feet firmly on the ground, by meeting a man who might fairly be called the founder of modern science: Galileo Galilei (1564–1642).

In Galileo's time there was a strong interest in the trajectories of cannonballs. It was, after all, a matter of life and death: an army that could judge how far gravity would allow a cannonball to fly would be better equipped to win a battle over a less well-informed enemy. Galileo's studies of the trajectory problem went far beyond those of any previous investigator. He made observations in the field and then performed careful experiments in the laboratory. These experiments are a model of care and attention to detail. He found out two things that startled many people in his day and that remain cornerstones of the science of gravity.

First, Galileo found that the rate at which a body falls does not depend upon its weight. Second, he measured the rate at which bodies fall and found that their acceleration is constant, independent of time.

After Galileo, gravity suddenly wasn't boring any more. Let's look at these two discoveries to find out why.

The story goes that Galileo took two iron balls, one much heavier than the other, to the top of the bell tower of Pisa and dropped them simultaneously. Most people of the day (and even many people today!) would probably have expected the heavier ball to have fallen much faster than the lighter one, but no: both balls reached the ground together.

The equality of the two balls' rates of fall went against the intuition and much of the common experience of the day. Doesn't a brick fall faster than a feather? Galileo pointed out that air resistance can't be neglected in the fall of a feather, and that to discover the properties of gravity alone we must experiment with dense bodies like stones or cannonballs, where the effects of air resistance are small. For such objects we find that speed is independent of weight.

But surely, one might object, we have to do much more work to lift a heavy stone than a light one, so doesn't this mean that a heavy stone "wants" to fall more than a light one and will do so faster, given the chance? No, said Galileo: weight has nothing to do with the speed of fall. We can prove that by measuring it. We have to accept the world the way we find it. This was the first step towards what we now call the *principle of equivalence*, which essentially asserts that gravity is indistinguishable from uniform acceleration. We shall see that this principle has a remarkable number of consequences, from the weightlessness of astronauts to the possibility of black holes.

Investigation 1.1. Faster and faster: the meaning of uniform acceleration

In this investigation, we work out what Galileo's law of constant acceleration means for the speed of a falling body. The calculation is short, and it introduces us to the way we will use some mathematical symbols through the rest of the book.

We shall denote time by the letter t and the speed of the falling body by v (for velocity). The speed at time t will be written $v(t)$. The acceleration of the body is g , and it is constant in time.

Suppose the body is dropped from rest at time $t = 0$. Then its initial speed is $v = 0$ at time $t = 0$, in other words $v(0) = 0$. What will be its speed a short time later?

Let us call this later time Δt . Here we meet an important new notation: the symbol Δ will always mean "a change in" whatever symbol follows it. Thus, a change in time is Δt . Similarly, we shall call the change in speed produced by gravity Δv . Normally we shall use this notation to denote small changes; here, for example, I have defined Δt to be "a short time later". We shall ask below how small Δt has to be in order to be "short".

The acceleration g is the change in the speed per unit time. This definition can be written algebraically as

$$g = \frac{\Delta v}{\Delta t}. \quad (1.1)$$

By multiplying through by the denominator of the fraction, we can solve for the change in speed:

$$\Delta v = g\Delta t. \quad (1.2)$$

Equation 1.1 basically defines g to be the *average* acceleration during the time Δt . If we take Δt to be very small, then this gives what we generally call the *instantaneous* acceleration. In this sense, "small" effectively means "as small as we can measure". If I have a clock which can reliably measure time accurate to a millisecond, then I would take Δt to be 1 ms if I wanted the instantaneous acceleration.

Now, Galileo tells us that the acceleration of a falling body does not in fact change with time. That means that the average acceleration during *any* period of time is the same as the instantaneous acceleration g . So in this particular case, it does not actually matter if Δt is small or not: Equation 1.1 is exactly true for any size of Δt . If we let t be any time, then we can rewrite Equation 1.2 as

$$\Delta v = gt.$$

We assumed above that the body was dropped from rest at time $t = 0$. This means that the initial speed is zero, and so the speed at a later time is just equal to Δv as given above. But if the body has an initial downward speed $v(0) = v_0$, then its subsequent acceleration only adds to the speed. This means that

$$v(t) = v(0) + \Delta v,$$

or

$$v(t) = gt + v_0. \quad (1.3)$$

Exercise 1.1.1: Speed of a falling body

Using the fact that the acceleration of gravity on Earth is $g = 9.8 \text{ m s}^{-2}$, calculate the speed a ball would have after falling for two seconds, if dropped from rest. Calculate its speed if it were thrown downwards with an initial speed of 10 m s^{-1} . Calculate its speed if it were initially thrown upwards with a speed of 10 m s^{-1} . Is it falling or still rising after 2 s?

The acceleration of gravity is uniform

Galileo performed a number of ingenious experiments with the rather crude clocks available in his day to demonstrate that the acceleration of falling objects is constant. Now, the acceleration of an object is the rate of change of its speed, so if the acceleration is constant then the speed changes at a constant rate; during any given single second of time, the speed increases by a fixed amount. We call this constant the **acceleration of gravity**, and denote it by g (for gravity). Its value is roughly 9.8 meters per second per second. The units, meters per second per second, should be understood as "(meters per second) per second", giving the amount of speed (meters per second) picked up per second. These units may be abbreviated as m/s/s , but it is more conventional (and avoids the ambiguous[†] ordering of division signs) to write them as m s^{-2} .

As with any physical law, there is no reason "why" the world had to be this way: the experiment might have shown that the speed increased uniformly with the distance fallen. But that is not how our world is made. What Galileo found was that speed increased uniformly with time of fall.

We can find out what Galileo's law says about the distance fallen by doing our first calculations, Investigation 1.1 and 1.2. These calculations show that uniform acceleration implies that the speed a falling body gains is proportional to time and that the distance it falls increases as the square of the time. The calculation also has another purpose: it introduces the basic ideas and notation that we will use in later investigations to construct computer calculations of more complicated phe-

In this section: near the Earth, bodies accelerate downwards at a uniform rate.

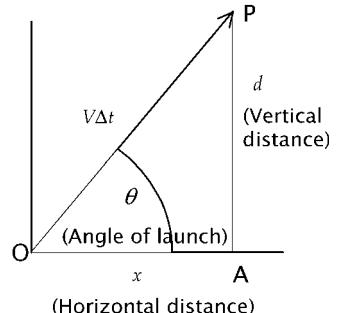


Figure 1.2. For the calculation in Investigation 1.3 on page 5, the vertical and horizontal distances traveled by a cannonball launched at an angle θ are the sides of a right triangle whose hypotenuse is the total distance $V\Delta t$.

[†]Ambiguity: does m/s/s mean $(\text{m/s})/\text{s}$ or $\text{m}/(\text{s/s})$? Either would be a valid interpretation of m/s/s , but in the second form the units for seconds cancel, which is not at all what is wanted.

Investigation 1.2. How the distance fallen grows with time

Here we shall calculate the distance $d(t)$ through which a falling body moves in the time t . Again we shall do this with simple algebra, but the ideas we use will lay the foundations for the computer programs we will write to solve harder problems later. Accordingly, much of the reasoning used below will be more general than is strictly necessary for the simple problem of a falling body.

We follow similar reasoning to that in Investigation 1.1 on the previous page. We are interested in the distance $d(t)$ fallen by the body by the time t . During the first small interval of time Δt , the body falls a distance Δd . (Our Δ notation again.) The *average* speed in this time is, therefore,

$$v_{avg} = \frac{\Delta d}{\Delta t}.$$

Solving this for Δd gives

$$\Delta d = v_{avg} \Delta t. \quad (1.4)$$

Now, we saw in Investigation 1.1 that the speed of the body changes during this interval of time. It starts out as zero (in the simplest case we considered) and increases to $g\Delta t$. So it seems to be an obvious guess that the average speed to use in Equation 1.4 is the average of these two numbers:

$$v_{avg} = \frac{1}{2}(g\Delta t + 0) = \frac{1}{2}g\Delta t.$$

If we put this into Equation 1.4, we find

$$\Delta d = (\frac{1}{2}g\Delta t)(\Delta t) = \frac{1}{2}g(\Delta t)^2. \quad (1.5)$$

I have said “an obvious guess” because it might not be right. If the acceleration of the body were a very complicated changing function of time, then its average speed over a time Δt might not be the average of its speeds at the beginning and end of the time-interval. For example, for some kinds of non-uniform acceleration it might happen that the body was at rest at the beginning and end of the interval, but not in between. Then its average speed might be positive, even though our guess would give zero.

Exercise 1.2.1: Distance fallen by a body

For the falling ball in Exercise 1.1.1 on the preceding page, calculate the distance the ball falls in each of the cases posed in that exercise.

Our guess is really only a good approximation in general if we choose the time-interval Δt small enough that the body's acceleration does not change by much during the interval. This gives a new insight into what is meant by a short time-interval: it must be short enough that the body's acceleration does not change by very much.

Of course, in the case of a falling body, the acceleration is constant, so we can expect Equation 1.5 to be *exact* for any time-interval, no matter how long. So if we replace Δt by t and Δd by $d(t)$, we find

$$d(t) = \frac{1}{2}gt^2. \quad (1.6)$$

Now suppose the body initially had a speed v_0 . Then the average speed during the time Δt would be $v_0 + g\Delta t/2$, so Equation 1.5 would become

$$\Delta d = (v_0 + \frac{1}{2}g\Delta t)\Delta t = v_0\Delta t + \frac{1}{2}g(\Delta t)^2.$$

Then, if the body does not start at distance $d = 0$ but rather at distance $d(0) = d_0$, we have that $d(t)$ at a later time is

$$d(t) = \frac{1}{2}gt^2 + v_0t + d_0. \quad (1.7)$$

This is the full law of distance for a uniformly accelerating body.

The calculation we have just done may seem long-winded, especially to readers who are comfortable with calculus, because the operations I have gone through may seem like a beginner's introduction to calculus. This is not my aim, however. It will become clear in future examples that what we have actually met here is a method of doing calculations by **finite differences**; this method is at the heart of most computer calculations of the predictions of physical laws, and we will see that it will help us to solve much more difficult problems involving the motion of bodies under the influence of gravity. We can use finite differences reliably provided we use intervals of time that are short enough that the acceleration of a body does not change by much during the interval.

nomena. Anyone who can do algebra can follow these investigations.

Trajectories of cannonballs

In this section: Galileo introduced the idea that the horizontal and vertical motions of a body can be treated separately: the vertical acceleration of gravity does not change the horizontal speed of a body.

We can now take up one of the subjects that contributed to the Renaissance interest in gravity, namely the motion of a cannonball. We have discovered that the vertical motion of the ball is governed by the law of constant acceleration. What about its horizontal motion? Here, too, Galileo had a fundamental insight. He argued that the two motions are *independent*.

Consider dropping a rubber ball in an airplane moving with a large horizontal speed. The rate at which the ball falls does not depend on how fast the plane is moving. Moreover, imagine an observer on the ground capable of watching the ball: it keeps moving horizontally at the same speed as the plane even though it is free of any horizontal forces. That is, while it falls “straight down” relative to the passengers in the plane, it falls in an arc relative to the observer on the ground.

Let us transfer this reasoning to the example of a cannonball launched at an angle to the vertical so that its vertical speed is v_0 and its horizontal speed is u_0 . Since there are no horizontal forces acting on the ball if we neglect air resistance, it will keep its horizontal speed as it climbs and falls, and the time it spends in the air will be the same as that of a ball launched vertically with the same speed v_0 . Galileo showed that the trajectory that results from this is a parabola.

This would be easy for us to show, as well, by doing a little algebra.

Investigation 1.3. The flight of the cannonball

Here we show how the finite-differences reasoning of the two previous investigations allows us to construct a computer program to calculate the flight of a cannonball, at least within the approximation that the ball is not affected by air resistance.

From this book's website you can download listing of the Java program CannonTrajectory. If you download the Triana software as well you can run the program and compute the trajectory of a cannonball fired at any given initial speed and at any angle. Figure 1.3 on the next page displays the result of the computer calculation for three trajectories, all launched with the same speed at three different angles. (The Triana software will produce plots of these trajectories. The figures produced for this book have, however, been produced by more sophisticated scientific graphics software.)

Here is how the program is designed. The idea is to calculate the body's horizontal and vertical position and speed at successive times spaced Δt apart. Let d be the vertical position and x the horizontal one, both zero to start. If the ball is launched with speed V at an angle θ with the horizontal (as in Figure 1.2 on page 3), then our first job is to deduce the vertical and horizontal speeds, which Galileo showed behaved independently of one another after launch.

Suppose that we turn off gravity for a moment and just watch a cannonball launched with speed V at an angle θ to the ground. Then after a small time Δt , it has moved a distance $V\Delta t$ in its launch direction. This is the distance OP in Figure 1.2. Simple trigonometry tells us that this distance is the **hypotenuse** of a right triangle whose other sides are the lines PA (the vertical distance d it has traveled) and AO (the horizontal distance x it has traveled). Then by definition we have

$$\begin{aligned}\sin \theta &= \frac{d}{V\Delta t} \Rightarrow d = V\Delta t \sin \theta, \\ \cos \theta &= \frac{x}{V\Delta t} \Rightarrow x = V\Delta t \cos \theta.\end{aligned}$$

The vertical speed is the vertical distance d divided by the time Δt , and similarly for the horizontal speed. We therefore find that the initial vertical speed is

$$v_0 = V \sin \theta$$

and the initial horizontal speed is

$$u_0 = V \cos \theta.$$

The horizontal speed remains fixed, so the horizontal distance increases by $u_0 \times \Delta t$ each time step. The vertical speed decreases by

$g\Delta t$ each time step, and we calculate the vertical distance using the *average* vertical speed in each time step. (In vertical motion, upward speeds are positive and downward ones negative.) The program sets up a **loop** to calculate the variables at successive time-steps separated by a small amount of time.

Normally one would expect a program like this to become more accurate for smaller time-steps, because of the remark we made in Investigation 1.2: our method of taking finite steps in time is better if the acceleration is nearly constant over a time step. In this case the relevant time step is Δt . By making Δt sufficiently small, one can always insure that the acceleration changes by very little during that time, and therefore that the accuracy of the program will increase. But in the present case that does not happen because our method of using the average speed over the time-step gives the *exact* result for uniform acceleration.

Let us look at the results of the three calculations in Figure 1.3 on the following page. Of these, the trajectory with the largest range for a given initial speed is the one that leaves the ground at a 45° angle. In fact it is not hard to show that this trajectory has the largest range of all possible ones. What is this range? We could calculate it from the results of Investigation 1.2, but in the spirit of our approach we shall try to guess it from the numerical calculation.

Given that the initial angle will be 45° , the range can only depend on the initial speed V and the acceleration g . The range is measured in meters, and the only combination of V and g that has the units of length is $V^2/g : (\text{m s}^{-1})^2/(\text{m s}^{-2}) = \text{m}$. We therefore can conclude that there is some constant number b for which $\text{range} = bV^2/g$. (This reasoning is an example of a powerful technique called **dimensional analysis**, because one is trying to learn as much as possible from the units, or **dimensions**, of the quantities involved in the problem.)

The numerical results let us determine b . Since the calculation used $V = 100 \text{ m s}^{-1}$, it follows that V^2/g is 1020 m. From the graph the range looks like 1020 m as well, as nearly as I can estimate it. Since the value of b is likely to be simple, it almost surely equals 1. An algebraic calculation shows this to be correct:

$$\text{maximum range} = V^2/g.$$

The reader is encouraged to re-run the program with various initial values of V to check this result.

Exercise 1.3.1: Small steps in speed and distance

Suppose that at the n^{th} time-step t_n , the vertical speed is v_n and the vertical distance above the ground is h_n . Show that at the next time-step $t_{n+1} = t_n + \Delta t$, the vertical speed is $v_{n+1} = v_n - g\Delta t$. Using our method of approximating the distance traveled by using the average speed over the interval, show that at the next time-step the height will be

$$h_{n+1} = h_n + \frac{1}{2}(v_n + v_{n+1})\Delta t = h_n + v_n\Delta t - \frac{1}{2}g(\Delta t)^2.$$

Exercise 1.3.2: Suicide shot

What is the *minimum* range of a cannonball fired with a given speed V , and at what angle should it be aimed in order to achieve this minimum?

Exercise 1.3.3: Maximum range by algebra

For readers interested in verifying the guess we made above from the numerical data, here is how to calculate the range at 45° algebraically. The range is limited by the amount of time the cannonball stays in the air. Fired at 45° with speed V , how long does it take to reach its maximum height, which is where its vertical speed goes to zero? Then how long does it take to return to the ground? What is the total time in the air? How far does it go horizontally during this time? This is the maximum range.

Exercise 1.3.4: Best angle of fire

Prove that 45° is the firing angle that gives the longest range by calculating the range for any angle and then finding what angle makes it a maximum. Use the same method as in Exercise 1.3.3.

But instead we show in Investigation 1.3 on the previous page how to use a personal computer to calculate the actual trajectory of a cannonball. These computer techniques will form the foundation of computer programs later in this book that will calculate other trajectories, such as planets around the Sun, stars in collision with one another, and particles falling into black holes.

Galileo: the first relativist

In this section: Galileo introduced what we now call the principle of relativity, which Einstein used as a cornerstone of his own revolutionary theories of motion and gravity almost 300 years later.

It would be hard to overstate Galileo's influence on science and therefore on the development of human society in general. He founded the science of **mechanics**; his experiments led the English scientist Isaac Newton (1642–1726) to discover his famous laws of motion, which provided the foundations for almost all of physics for 200 years. And almost 300 years after his death his influence was just as strong on Albert Einstein. The German–Swiss physicist Einstein (1879–1955) replaced Newton's laws of motion and of gravity with new ones, based on his theory of relativity. Einstein's revolutionary theories led to black holes, the Big Bang, and many other profound predictions that we will study in the course of this book. Yet Einstein, too, kept remarkably close to Galileo's vision.

The main reason for Galileo's influence on Einstein is that he gave us the first version of what we now call the **principle of relativity**. We have already encountered Galileo's version: the vertical motion of a ball does not depend on its horizontal speed, and its horizontal speed will not change unless a horizontal force is applied.

Where we used a fast-flying airplane to justify this, Galileo imagined a sailing ship on a smooth sea, but the conclusion was the same: an experimenter moving horizontally will measure the same acceleration of gravity in the vertical direction as he would if he were at rest.

Galileo took this idea and drew a much more profound conclusion from it. The radical proposal made half a century earlier by the Polish priest and astronomer Nicolas Copernicus (1473–1543), that the Earth and other planets actually moved around the Sun (see Figure 1.4), was still far from being accepted by most intellectuals in Galileo's time. Although the proposal explained the apparent motions of the planets in a simple way, it was open to an important objection: if the Earth is moving at such a rapid rate, why don't we *feel it*? Why

isn't the air left behind, why doesn't a ball thrown vertically fall behind the moving Earth?

Galileo used the independence of different motions to dispose of this objection. Galileo's answer is that a traveler in the cabin of a ship on a smooth sea also does not feel his ship's motion: all the objects in the cabin move along with it at constant speed, even if they are just resting on a table and not tied down. Anything that falls will fall vertically in the cabin, giving no hint of the ship's speed. So it is on the Earth, according to Galileo: the air, clouds, birds, trees, and all other objects all have the same speed, and this motion continues until something interferes with it. There is, in other words, no way to tell that the Earth is moving through space except to look at things far away, like the stars, and see that it is.

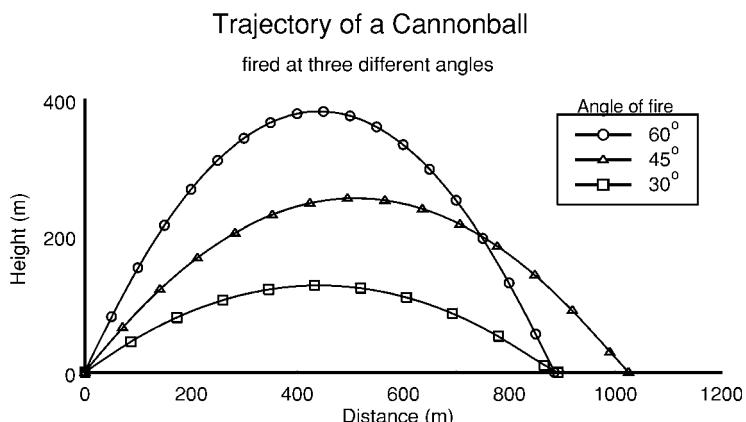


Figure 1.3. Trajectories computed by the program developed in Investigation 1.3 on the previous page, for three angles of firing, each at the same initial speed of 100 m s^{-1} . The trajectory at 45° goes furthest.

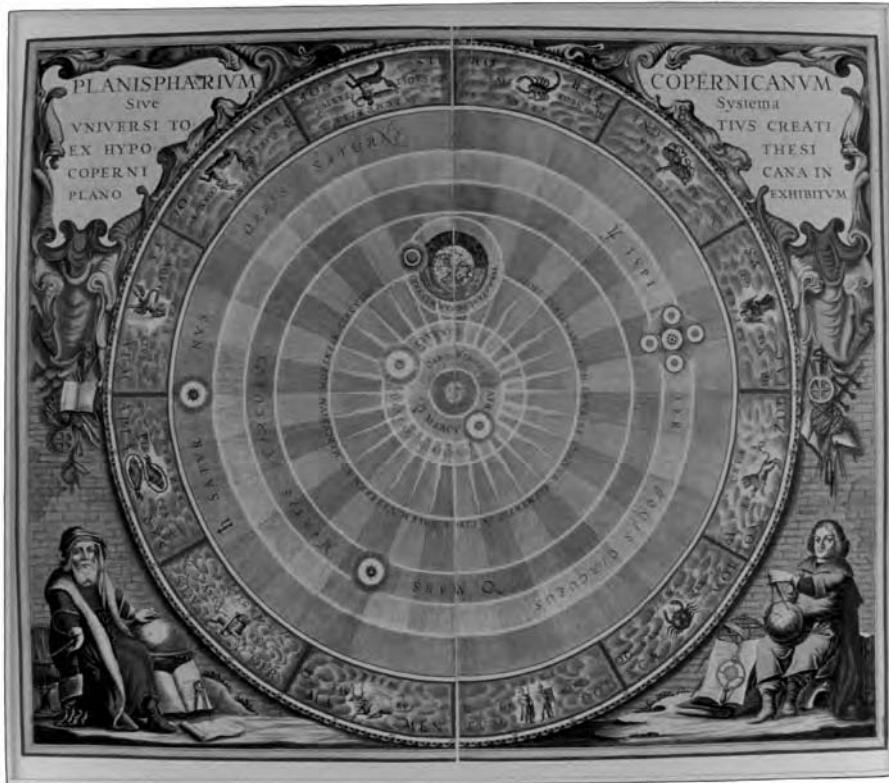


Figure 1.4. The Copernican view of the planets known in Galileo's day as they orbit the Sun.

Today we re-phrase and enlarge this idea to say that *all* the laws of physics are just the same to an experimenter who moves with a uniform motion in a straight line as they are to one who remains at rest, and we call this the *principle of relativity*. We shall encounter many of its consequences as we explore more of the faces of gravity.

Unfortunately for Galileo, his clear reasoning and his observations with one of the first telescopes made him so dangerous to the established view of the Roman Catholic Church that in his old age he was punished for his views, and forced to deny them publicly. Privately he continued to believe that the planets went around the Sun, because he had discovered with his telescope that the moons of Jupiter orbit Jupiter in the same way that the planets orbit the Sun.

Today we recognize Galileo as the person who, more than anyone else, established the Copernican picture of the Solar System.

1
2	o
3
4
5
6
7
8
9
10

Figure 1.5. Part of a sketch by Galileo of the positions of Jupiter (open circles) and its moons (stars) on a sequence of nights (dates given by the numbers). The big changes from night to night puzzled Galileo. At first he believed that Jupiter itself was moving erratically, but after a few observations he realized that the "stars" were moons orbiting Jupiter in the same way that the planets orbit the Sun.

And then came Newton: gravity takes center stage

Born in the same year, 1642, as Galileo died, Isaac Newton revolutionized the study of what we now call physics. Part of his importance comes from the wide range of subjects in which he made fundamental advances – mechanics (the study of motion), optics, astronomy, mathematics (he invented calculus), . . . – and part from his ability to put physical laws into mathematical form and, if necessary, to invent the mathematics he required. Although other brilliant thinkers made key contributions in his day – most notably the German scientist Gottfried Leibniz (1646–1716), who independently invented calculus – no physicist living between Galileo and Einstein rivals Newton's impact on the study of the natural world.

Nevertheless, it is hard to imagine that Newton could have made such progress in the study of motion and gravity if he had not had Galileo before him. Newton proposed three fundamental **laws of motion**. The first two are developed from ideas of Galileo that we have already looked at:

The *first law* is that, once a body is set in motion, it will remain moving at constant speed in a straight line unless a force acts on it. This is just like the rubber ball dropped inside the airplane of Chapter 1.

This is basically Galileo's idea, which led him to his principle of relativity, that motions in different directions could be treated independently. Notice that, since particles travel in *straight* lines unless disturbed, the directions along which motion is independent must also be along straight lines.

Newton's *second law* is that, when a force is applied to a body, the resulting *acceleration* depends only on the force and on the mass of the body: the larger the force, the larger the acceleration; and the larger the mass of a body, the smaller its acceleration.

This dependence of the acceleration a on the force F and the mass m can be written as an equation:

$$a = F/m.$$

It is more conventional to write it in the equivalent form

$$F = ma. \quad (2.1)$$

The second law: weight and mass

The second law fits our everyday experience of what happens when we push something. If we have a heavy object on wheels (to allow us to ignore friction for a moment) and we give it a push, its speed increases (it accelerates) as long as we continue to push it. Then it moves along at a constant speed after we stop pushing. (Friction eventually slows it down, but that is just another force exerted by the surface it is moving across.) If we push it harder, it accelerates faster, so it is not

In this chapter: we learn about Newton's postulate, that a single law of gravity, in which all bodies attract all others, could explain all the planetary motions known in Newton's day. We also learn about Newton's systematic explanation of the relationship between force and motion. When we couple this with Galileo's equivalence principle, we learn how gravity makes time slow down.



Figure 2.1. Brilliant and demanding, Isaac Newton created theoretical physics. Besides devising the laws of gravity and mechanics, he invented calculus, still the central mathematical tool of physicists today. (Original engraving by unknown artist, courtesy AIP Emilio Segrè Visual Archives, Physics Today Collection.)

In this section: how force, mass, and acceleration are related to one another, and the difference between weight and mass.

unreasonable to guess that the acceleration might be proportional to the force we exert on it. Moreover, if we load more things on top of the object we are pushing, then to get the same acceleration, we need to push harder, so again we might guess that the force required would be proportional to the mass. Newton not only made these guesses, but he assumed (or hoped!) that the force would depend on nothing else besides the mass of the body and the acceleration produced.

►A good illustration of how everyday language uses such terms differently from the way we use them in physics is provided by dieting. About twice a year, when I go on a diet, I tell my friends that I am trying to lose weight. Mercifully, none of them has yet pointed out that the surest way to lose weight is to go to the Moon! That would not really help, of course, since what I am really trying to do, for the sake of my health, is to lose *mass*. If I stay on the Earth, then of course losing weight implies losing mass.

Newton made an important distinction between two concepts that are often used interchangeably in everyday language: **mass** and **weight**. The mass of an object is, as we have just seen, the way it “resists” being accelerated. (Physicists sometimes call this its **inertia**.) The weight of an object is the force of gravity on it. When we step on our bathroom scales, we measure our weight; if we were to put the scales on the Moon, where gravity is weaker, we would get a lower reading. Our mass would not have changed, however. It would take the same force to accelerate us on a smooth horizontal track on the Moon as it would on the Earth.

The second law leads to a remarkable insight when combined with Galileo’s discovery that bodies of different masses accelerate under gravity at the same rate: it tells us that a body’s *weight* must be proportional to its *mass*. Here is the reasoning.

Suppose we lift a heavy body off the floor and hold it. What we *feel* as its weight is really the sensation of exerting an upwards force upon it to hold it against the force of gravity. When we exert this force on a body, it remains at rest in our hands. By the first law, we conclude that the total force on it is zero: our upwards force just cancels the downwards force of gravity on the object. Therefore the weight of the body *is* the force of gravity on it. If we now release the body, the force of gravity on it is the same but is no longer balanced by our hands’ force. The body accelerates downwards: it falls.

But what *is* its acceleration in response to this force of gravity? Galileo observed that the acceleration of the body does not depend on its weight: it is the *same* for everything. Now, in Equation 2.1 on the previous page, the only way that we can change the force F (the weight) without changing the acceleration a is if we change the mass m in proportion to F : the force of gravity on a body is proportional to its mass.

Newton’s reasoning here leads to an experimentally verifiable conclusion. Both mass and weight could be measured independently, the weight using scales and the mass by measuring the acceleration of the body in response to a given *horizontal* force and then using Equation 2.1 on the preceding page to infer its mass. If we divide the weight by the mass, we should get the acceleration of gravity. Put mathematically, this says that the force of gravity F_{grav} on any object equals its mass m times the acceleration of gravity g :

$$F_{\text{grav}} = mg. \quad (2.2)$$

In honor of Newton, scientists have agreed to measure force in units called *newtons*: one newton (N) of force equals 1 kg times an acceleration of 1 m s⁻². The weight of a body in newtons is then just its mass in kilograms times the acceleration of gravity, 9.8 m s⁻².

As we have noted, Equation 2.2 is experimentally verifiable: if it holds for any body, then this experiment serves as a test of the second law itself.

Once this law of motion was checked experimentally, Newton’s argument led to a reformulation of Galileo’s *principle of equivalence*: the mass of a body (ratio of force to acceleration) is proportional to its weight. From this statement and Equation 2.1 on the preceding page,

Galileo's original observation that all bodies fall with the same acceleration follows.

This was the way two centuries of physicists thought of the principle of equivalence. It was strikingly confirmed by the Hungarian physicist Baron Roland von Eötvös (1848–1919) in 1889 and in 1908. In one of the most accurate physics experiments of his time, von Eötvös showed that many different materials fell with the same acceleration, to within a few parts in a billion!

We now know that this form of the equivalence principle applies not just to ordinary bodies, but also to bodies with the strongest of gravitational fields in general relativity, even to black holes. However, Einstein's general theory of relativity did change Newtonian mechanics in some respects, so the way that modern physicists think about the principle of equivalence is also rather different from the Newtonian form. We will have to wait a few pages before we take a look at the modern reformulation, because we have not by any means finished with Newton's work yet. He made two further landmark contributions to our subject: his third law of motion and his law of gravitation.

The third law, and its loophole

Newton added another law, not explicit in Galileo's work, but which he needed to make the whole science of mechanics self-consistent.

Newton's *third law* states that if I exert a force on an object, then it exerts a force back on me that is exactly equal in magnitude and opposite in direction to the one I have applied. This law is often paraphrased as "action equals reaction".

This law often strikes newcomers to the subject as contradictory: if there are two equal and opposite forces, don't they cancel? If the object I push on moves and I don't, doesn't that mean that I pushed harder on it than it pushed on me? These difficulties are always the result of mis-applying Newton's second law. Only the forces acting *on* an object contribute to its acceleration. The equal and opposite forces in the third law act on different objects, and so there is no way that they can cancel each other.

To see our way past such doubts with a concrete example, consider again the feeling of a weight in the hand. Suppose I hold an apple. I have to exert a force on the apple, equal to its weight, to keep it in one place. This force is exerted through my hand, but the hand doesn't stay where it is all by itself: it is kept there by the force exerted on it by my arm. (The tired feeling I eventually get in the muscles of my arm leaves me no room to doubt this!) Since the hand isn't moving but the arm is exerting a force on it, there must be a balancing force on it as well, and this can only come from the apple. So as I exert a force on the apple, it exerts a force back on me. How much of a force? Newton argued that this "reaction" by the apple must be equal to the force I exert on it.

There are several ways of seeing that this is reasonable. Suppose, for instance, that the force exerted by the apple on my hand was only half of its weight. What makes the hand special, that it gets back only half the force it gives out? Why wouldn't it be the other way around, that the apple should receive from my hand only half the force it exerts on my hand? This lack of symmetry, where the hand gets only half the force back that it exerts, while the apple gets twice its force back, makes the "half-reaction" law illogical.

The third law has an important practical consequence: it is responsible for almost all propulsion. We walk by pushing backwards with our feet

In this section: when you push on a body, it pushes back on you. Normally the two forces are the same size. But when bodies are separated, they can differ, and this leads to a force called radiation reaction.

on the ground; the ground then pushes forwards on us, and that is why we move. Similarly, a rocket pushes hot gas out of its nozzles in the backwards direction; the gas exerts its reaction to push the rocket forwards. Jet planes, swimming fish, and flying birds all use the third law to get around. What about sailing ships and downhill skiers?

Logical as the third law is, there is a loophole in it which we will find important when we discuss general relativity. The argument we have just gone through applies strictly only to forces exerted by bodies in constant contact; that is, by bodies that are touching and not moving relative to each other. But some bodies can exert forces without being in contact.

Electric charges, for example, exert forces over considerable distances, and these forces weaken as the distance increases. Now, the laws governing electric and magnetic forces (which together are called **electromagnetism**) tell us that these forces can only be transmitted at the speed of light, no faster. For example, if two electric charges move relative to one another, then the forces they exert on one another have to change as the distance between them changes. These changes travel through space at the speed of light. It thus takes some time for the force to be transmitted over a distance, so there will be a delay between the time when charge "A" exerts its force on charge "B" and the time when it receives the reaction force from "B". The electric charges will have changed their relative distances in this time, and then the two forces will not be equal.

In fact, this imbalance of forces leads to what is called **radiation reaction** on moving, accelerating charges: an accelerating charge experiences a force that depends on the rate of change of its acceleration. This force usually opposes its motion and acts as a kind of friction. When we send out radio waves from a radio transmitter, we have to drive the electrons in the transmitter into exactly the right sort of motions to produce the desired radio waves. The power required to transmit from such antennas goes primarily into overcoming the radiation reaction force on the charges. Newton knew nothing about this and did not allow for it in his formulation of the third law. We shall return to the subject of radiation reaction in more detail in Chapter 22 when we discuss gravitational radiation.

We would therefore be safer paraphrasing the third law as: "action equals reaction for bodies in constant contact".

Preview: Newton's gravity

In this section: with a leap of imagination that even Galileo had not attempted, Newton showed that the same gravity that made apples fall to Earth makes the Moon stay near the Earth and the Earth near the Sun. A simple mathematical formula was consistent with all known data for planetary orbits.

Now we come to another of Newton's jewels, his law for the gravitational attraction between the Sun and its planets. We will look much more closely at its consequences when we study planetary motion in Chapter 4, but we describe it briefly here, not only because it is closely linked to his laws of motion and the equivalence principle, but also because in the law of gravity we see Newton's imagination at its boldest.

Believing that the laws of motion ought to apply everywhere, not just on the surface of the Earth, Newton realized that the orbital motion of the Moon about the Earth and of the planets about the Sun implied that the heavenly bodies were under the influence of *forces*, forces which moreover had to be exerted over considerable distances, since telescopes didn't reveal any horse carts pushing the planets around. (Greek and Roman mythology didn't stand up to the test of observation!)

The Earth already exerted the force of gravity on objects (falling apples, etc.) that were not in direct contact with it, so could the Earth's gravity extend very much farther away? Could gravity be the force responsible for keeping the Moon in its orbit?

To answer “yes” to this question, something we take so much for granted today, was in Newton’s day a brilliant and even courageous leap. Newton realized that he would be ignored and even ridiculed unless he could extend the law of gravity to the heavens in a simple and convincing way. His courage was rewarded: in Galileo’s time the German astronomer Johannes Kepler (1571–1630) had shown from painstaking calculations that the planets followed elliptical orbits with the Sun always at one focus of the ellipse, and Newton was able to find a force law that predicted exactly that.

The law of gravity *was* simple and, in the end, utterly convincing: the gravitational force between two bodies is proportional to the product of their masses and inversely proportional to the square of the distance between them.

This can be expressed as an equation. If the two bodies have masses M_1 and M_2 , respectively, and are separated by a distance r , then the force of gravity exerted by each on the other is

$$F_{\text{grav}} = \frac{GM_1M_2}{r^2}, \quad (2.3)$$

where G is a constant of proportionality, called *Newton’s gravitational constant*. Its value, as measured by the best modern experiments, is $6.6720 \times 10^{-11} \text{ m}^3 \text{ s}^{-2} \text{ kg}^{-1}$. The dependence of the force on the masses M_1 and M_2 of the bodies follows directly from our discussion of the equivalence principle: the force of gravity on a body is its weight, and this must be proportional to its mass. By the third law, the other body will experience the same force, so the force must be proportional to its mass, too. The only new element in this law is the way it depends on distance r . We call it an **inverse-square** law.

How Newton decided that the dependence on r should be as $1/r^2$ is a story that shows that Newton had more than just courage: he had the capacity for immense hard work and persistence. For in order to prove to himself that the inverse-square-law force gave elliptical orbits, he had to invent the calculus. And even after proving the law to himself, he still refrained from publishing it for many years until he could iron out a particularly difficult detail that he felt might otherwise have proved his undoing. We will return to this difficulty when we examine the orbits predicted by this force, in Chapter 4. Newton’s delay in publishing his work also led to bitter disputes with some contemporaries, especially with Robert Hooke (1635–1703), who claimed to have invented the inverse-square law himself.

Action at a distance

Newton’s gravitational force between heavenly bodies was an *instantaneous* force: no matter how far apart two bodies were, the force between them would respond instantly to any change in their separation. Newton did not try to invent any mechanical way of describing the force, say by hypothetical particles traveling between one body and another – such intermediaries would travel at a finite speed, not instantly.

The way Newton’s gravitational force behaved was called **action at a distance**. It was one of the most unpalatable aspects of his theory to many of his contemporaries, and Newton was forced to defend it vigorously. One can see why he needed it, for otherwise he could not have argued that his third law of motion (action-reaction) would apply in the heavens: the argument we gave above – that forces that are transmitted with a delay don’t necessarily have equality of action and reaction – would have been devastating. In order to preserve his third law, and (very importantly)



Figure 2.2. Johannes Kepler was the foremost mathematical astronomer of his time. Working in Prague, his detailed studies of the motions of the planets made it possible for him to show that orbits were ellipses long before Newton explained this. The calculation was immense: more than a thousand sheets of arithmetic for his calculation of the orbit of Mars survive. Kepler was also an accomplished mathematician, proving the close packing theorem for spheres and explaining why logarithms worked. Image by permission of AIP Emilio Segrè Visual Archives.

In this section: Newton’s law was audacious: gravity could act across distances apparently without anything in between. Einstein removed that and returned to a model for gravity that more resembles other influences: gravitational effects move through space at the speed of light. But Newton’s law was the right one for his time, a single simple assumption that could explain a wealth of data.

in the absence of any evidence to the contrary, Newton held firm to instantaneous forces.

Einstein's general relativity replaced Newton's law of gravity with a more complicated theory that does have a finite speed of transmission of gravitational influences (the speed of light), so modern gravity does not involve action at a distance. But it would be hard to find a physicist today who would argue that Newton was "wrong" to take the position he did. Newton's sense of how to make progress with gravitational theory was unerring: by keeping it simple he provided physicists with a law which satisfactorily explained Solar System motions for two hundred years. And when some tiny discrepancies with his theory were finally observed (in details of the orbit of Mercury), the instantaneous aspect of his theory was not to blame. We will see in Chapter 21 that these discrepancies were finally explained by the curvature of space in general relativity.

Newton's theory of gravity did not immediately affect the study of gravity on the Earth, but in common with his three laws of mechanics it shows how he relied on a sense of simplicity to formulate his physical laws. The world is a complicated place, but within its complexity Newton tried to formulate his laws as simply as experience would allow. The third law, relying on symmetry between bodies, is an example: there was no experimental hint of the exceptional case we pointed out above (bodies in relative motion), so he did not allow for it.

Newton's law of gravity was likewise much simpler than the data he was trying to explain. It had a previously unknown constant of proportionality in it (G), and a perhaps surprising exponent (*inverse-square* law), but with just those two numbers he could explain the huge number of detailed observations of the Moon and the five known planets. Moreover, by uniting the theories of planetary motion and of terrestrial gravity, we can justly give him credit for having devised the first **unified field theory**. The theme of simplicity in the face of complex phenomena, especially of simplification by unification, has ever since been a dominant one in physicists' attempts to explain the world. We shall see that it applies particularly to Einstein's relativity, despite its exaggerated reputation for difficulty.

The new equivalence principle

In this section: Galileo's equivalence principle is reformulated into its modern version. An experimenter who falls freely in a gravitational field measures no gravity at all nearby. Weightlessness of astronauts is the classic example.

As I mentioned above, the modern view of the equivalence principle is somewhat different from both Galileo's and Newton's. This is because experiments have taught us that light has some special properties that make it hard to fit into Newtonian gravity. Newton couldn't have known this, for it was not possible to study light accurately with the technology of his day. But by the late nineteenth century, experiments began to force physicists to realize that light was somehow special. Today we know that it is fundamental to Einstein's **special relativity** theory that light has a *fixed* speed, called c , about 3×10^8 m s⁻¹. (I shall deliberately leave this statement a little vague here, and explain what a "fixed speed" means in Chapter 15.) We have also learned that light has *zero* inertial mass.

It is clear that the motion of light will not be described by the simple law $F = ma$, and since our formulation of the equivalence principle involves this equation, we might feel that light could violate the principle. However, since even the meaning of acceleration is unclear in the case of light, it is more sensible to reformulate the equivalence principle without mentioning acceleration or inertial mass. We will then see in the next section that, once we have done this, we can actually make striking predictions about the effect of gravity on light, without ever referring to $F = ma$.

One thing Galileo might have observed, had he been inclined to think this way about the problem, was that if he himself had fallen off the Leaning Tower of Pisa

Investigation 2.1. The effect of motion on light: the Doppler effect

One of the most important effects in the study of light, or indeed of any wave, is the Doppler effect. For sound waves, the Doppler effect causes the whistle of an approaching train to sound at a higher pitch than if the train were at rest. And as the train moves away, the pitch falls to a lower note. Analogous things happen to light.

In outer space, far from any gravitational fields, light travels in a straight line at the constant speed c and with unchanging frequency (color). If two experimenters moving relative to one another look at the same beam of light, they will generally see it to have different colors: this is called the Doppler shift of light. In particular, if one experimenter measures the light to have frequency f_0 , and if for simplicity the second experimenter is moving in the same direction as the light is going, only with a speed v that is small compared to c , then the second experimenter will measure the frequency to be

$$f_1 = f_0(1 - v/c). \quad (2.4)$$

This is smaller than f_0 , and it means that, for example, blue light will be shifted toward the red end of the spectrum, and other colors will be shifted in the same sense. This is therefore called a *redshift* of light. In Figure 2.3 there is a visual derivation of the Doppler effect, leading to Equation 2.4.

If, on the other hand, the second experimenter were moving directly towards the source of the light, then we could use Equation 2.4 with v replaced by $-v$. This would increase the frequency of the light and result in a *blueshift*.

The important lesson to understand is that the frequency of a light beam is not an intrinsic property of the beam; rather it depends also on the motion of the experimenter who measures it. When we discuss the redshift of light produced by a gravitational field, we shall have to be careful to define who the experimenter is who measures the redshift, since other experimenters could see the same light with a blueshift.

along with the two heavy balls, then on the way down the balls would simply have stayed beside him: they would have behaved relative to him as if they had had no forces on them at all. (Again we are neglecting air resistance here, for clarity.)

Precisely, this means that if he had given one of the balls a push in any direction, even downwards, its subsequent motion relative to him would have been with uniform speed in a straight line. Both he and the ball would have been accelerating downwards relative to the Earth, but when we talk about their motion relative to one another, this common acceleration subtracts away exactly, leaving only uniform relative motion.

This happens only because, for gravity, the acceleration of every body is the same. Such statements would not be true of other forces, like electromagnetism. This lack of relative acceleration, even in a gravitational field, allows us the following formulation of the equivalence principle.

In a gravitational field, all objects behave in such a manner that they appear to be completely free of any gravitational forces when observed by a freely-falling experimenter.

To a freely-falling Galileo, the laws of physics would be the same as they would be in outer space, far from any gravitating bodies. This is a formulation that Galileo and Newton would have accepted, even if they would have found it strange, and it is particularly suitable for us because all mention of mass and acceleration has disappeared from it.

It is possible to perform a home experiment that directly illustrates this version of the equivalence principle very well. It is a "toy" that was given to Einstein on his 76th birthday, described in Figure 2.5 on page 17.

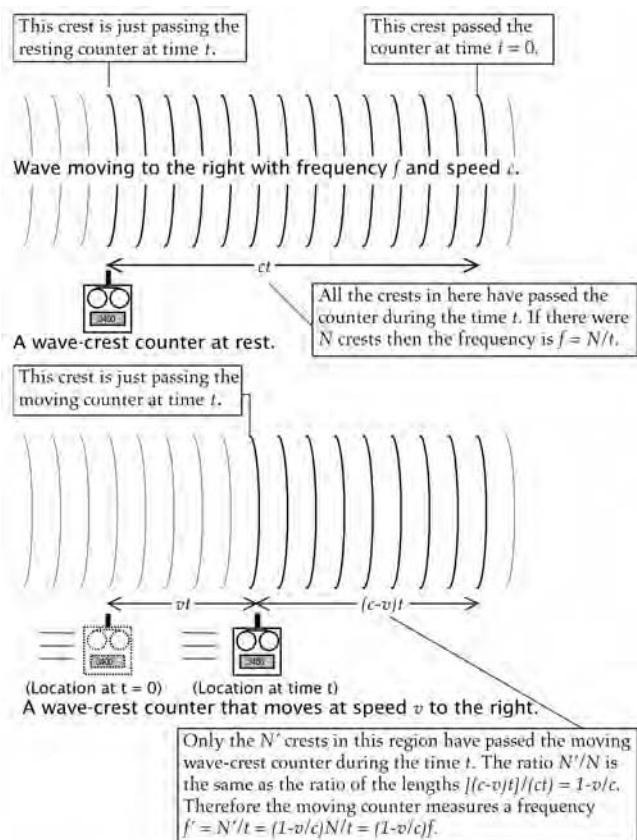


Figure 2.3. A visual derivation of the Doppler effect. The wave runs from left to right. The counter counts the number of crests of the wave that pass it. The frequency of the wave is the number of crests per unit time. This depends on whether the counter is moving or not.

Investigation 2.2. The effect of gravity on light: the redshift

To understand how gravity affects light, we imagine a beam of light of a particular frequency f_{bottom} that is shining upwards from a source on the ground. An experimenter stands on a tower of height h directly over the source, and he measures the frequency of the light when it reaches him. He calls this frequency f_{top} . What is the relation between f_{bottom} and f_{top} ? We will use the equivalence principle, which means introducing another experimenter who is freely-falling.

Suppose, then, that the experimenter at the top has a companion who falls off the tower at the moment that the beam of light leaves the ground. (Fortunately this is only a thought experiment!) As a freely-falling experimenter, the companion finds that light moves as if it were in outer space; in particular, the frequency that he measures does not change with time. At the instant he leaps off, he is still at rest with respect to the ground, so he measures the same frequency f_{bottom} as an experimenter on the ground would measure. By the equivalence principle, this is also the frequency he (the companion) measures when the light reaches the top of the tower a moment later.

But in this brief moment, the companion has begun to fall. Relative to the companion, the experimenter at the top is moving away from the light source at the time of reception of the light at the top, and so the fixed experimenter's frequency f_{top} will be *redshifted* with respect to the companion's frequency f_{bottom} . If our version of the equivalence principle is right, then light is redshifted as it climbs out of a gravitational field.

Exercise 2.2.1: Redshift to a satellite

Calculate the redshift gh/c^2 if h is the distance from the ground to a satellite in low-Earth orbit, 300 km. Suppose the "light" is actually a radio wave with a frequency of 10^{11} Hz. How many cycles would the transmitter emit if it ran for one day? How many fewer would be received in one day by the satellite? How long did it take the transmitter to generate these "extra" cycles?

We can find out how much this redshift is by calculating the speed of the falling companion when the light arrives at the top. To travel a distance h , the light takes a time h/c . In this time the companion falls with the acceleration of gravity, g . His final speed v is therefore just g times the time of fall, or gh/c . From Equation 2.4 on the previous page we then have

$$f_{\text{top}} = f_{\text{bottom}}(1 - gh/c^2). \quad (2.5)$$

The magnitude of the effect is very small, but not too small to be measured. For example, if the tower were 100 m high, then in Equation 2.5 gh/c^2 would be only 1.1×10^{-14} , so the change in the frequency of light would have to be measured to an accuracy of a few parts per 10^{15} . Two very high precision experiments by R V Pound, G A Rebka, and J L Snider in the 1960s confirmed the effect with good accuracy. (See Figure 2.4.) Today it is checked every day when routine corrections for the redshift are put into the time-signals of the GPS satellite system, as described in the text.

It is important to understand that, as we remarked at the end of Investigation 2.1 on the preceding page, the redshift is a property of the experimenters as well as of light. Thus, light at any height does not have a unique frequency. If we measure its frequency using experimenters at rest relative to the ground, then there will be a redshift. If we measure its frequency by using freely-falling experimenters, there will be none.

The gravitational redshift of light

In this section: a direct consequence of the equivalence principle is that light changes its frequency as it climbs out of a gravitational field. It shifts toward the red end of the spectrum.

The effect of gravity on light can now be found by demanding that it should behave as if there were no gravity when it is observed by a freely-falling experimenter. This means, in particular, that it should follow a straight line with no change in its frequency. In Investigation 2.1 on the previous page we show that the frequency of light is affected by the motion of an experimenter. This is called the *Doppler effect*. Light can experience a **redshift** or a **blueshift**, depending on the motion of the source relative to the experimenter. Then in Investigation 2.2 we show how the Doppler effect and the new equivalence principle combine to tell us how gravity affects light as it moves upwards from the ground.

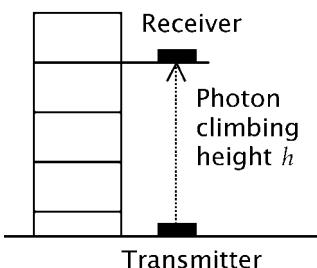


Figure 2.4. A sketch of the experimental arrangements for the Pound-Rebka-Snider experiment, which first detected the gravitational redshift on Earth.

The result is that the frequency of a beam of light climbing up out of a gravitational field is Doppler-shifted towards lower frequencies. Blue light becomes more red, so this is called the *gravitational redshift*.

It is important to remember that the frequency of light depends on the state of motion of the experimenter who measures it, so that when we say that light suffers a gravitational redshift, we mean that the experimenters must both be at rest relative to the Earth; a freely-falling experimenter, for example, should see no redshift at all.

When general relativity was first proposed, physicists soon realized that it predicted the gravitational redshift of light, and observations of the redshift of light leaving the Sun and other stars were regarded as an important test of the validity of general relativity. But we can see from our discussion that *any* theory of gravity can be expected to predict the effect if it respects the principle of equivalence. The modern view is that observations of the gravitational redshift test the equivalence principle.

Gravity slows time

The gravitational redshift leads us to a very profound conclusion about time itself: gravity makes it run slower. Suppose we build two identical clocks, in such a way that the clocks tick once in every period of the oscillation of the electromagnetic wave that we use in the redshift experiment. We place clock one on top of the tower in the experiment and leave the other on the ground. By design, the one on the ground ticks at the same rate as the frequency of the light signal that we emit there. Suppose we keep it there for, say, 10^{20} ticks. (Since visible light oscillates at about 10^{15} times a second, this would be about one day.) Now, the clock at the top of the tower receives the light redshifted, so the light frequency at the top of the tower is less by one part in 10^{14} (see Equation 2.5). The clock at the top is therefore ticking faster than the arriving light by that same factor.

Now, the light going up the tower is just a wave; one oscillation corresponds to the arrival of one “crest” of the wave. Crests don’t disappear on the way up, so exactly as many oscillations of light arrive at the top during the experiment as were emitted at the bottom: 10^{20} in this case. But during the experiment, the clock at the top has ticked more times, by one part in 10^{14} . That means it has ticked 10^6 times more than the one on the ground. When the experiment finishes, we immediately bring the clock from the top of the tower down to the ground and compare it to the one on the ground. The one that has been sitting on the tower is ahead of the one on the ground, by these 10^6 ticks. This is only one nanosecond ($1 \text{ ns} = 10^{-9} \text{ s}$), but it is measurable.

Let us take stock of what we have learned. Given two identical clocks, if we place one for a while higher up in a gravitational field and then bring it down to the other one, we will find it has gone faster. This conclusion applies to any clock, biological or physical, regardless of how it is made: the workings of the clock did not come into the argument above.

Since all clocks run faster higher up, we conclude that time itself runs faster higher up in the gravitational field: after all, time is only what we measure using clocks.

This is not just an abstract point. Today there are in orbit around the Earth a number of satellites that form the Global Positioning System (GPS). Launched by the US Air Force, they constantly send radio signals down to Earth that can be used in navigation: with a GPS receiver one can pinpoint one’s location to within 10 m, an extraordinary accuracy. The satellites carry precise atomic clocks, the most accurate clocks that can be made. Because of the effect of gravity on time, these tick faster than do clocks on the ground; the difference is about three microseconds per day. (A microsecond or μs is 10^{-6} s .) Yet to give a position that is accurate to 10 m requires clocks that are accurate to the time it takes the radio waves from the satellites to travel 10 m, which is $0.03 \mu\text{s}$. Therefore, this redshift correction *must* be taken into account in order for the system to function. (Actually, there is also a velocity correction that we will go into in Chapter 15, and this has to be taken into account as well.)

The routine use of the GPS by airplanes, ships, long-distance trucks, and even private cars confirms the gravitational redshift and the effect of gravity on time to a much higher accuracy than the original Pound–Rebka–Snider experiment.

In this section: from the redshift of light it follows that time itself slows down when gravity is strong. The GPS navigational satellite system, which relies on highly accurate clocks, must take this effect into account in order to maintain its accuracy.

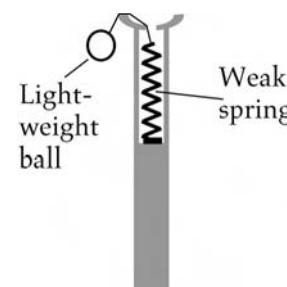


Figure 2.5. Einstein’s equivalence-principle toy. This “toy” was a gift to Einstein from E M Rogers. It consists of a light ball tied to a spring too weak to pull the ball into the cup. The secret of getting the ball into the cup is to make it weightless. Then the spring will be able to pull it in. Your challenge is, how do you make the ball weightless? Einstein delighted in demonstrating the equivalence principle with this toy. The toy and Einstein’s reaction are described in Einstein: A Centenary Volume, ed. A P French (Heinemann 1979), pp 131–132.

Summing up

In this section: the seemingly simple consideration of gravity on

Earth has given us the tools we need to study most of the Universe that modern astronomy reveals to us. But, in addition, it has brought us to deep conclusions: the effect of

gravity on the color of light, the slowing of time by gravity. These ideas form the foundation on which we will build the modern theory of gravity later in this book.

Although we have confined ourselves in these first two chapters mainly to gravity on the Earth's surface, and we have explored its properties with simple experiments and straightforward reasoning, we have uncovered a rich treasure of different ideas. We have been led to the principle of equivalence, the principle of relativity, Newton's laws of motion, the gravitational redshift of light, and the fact that gravity affects time itself.

We also now have a more complete answer to Copernicus' critics than Galileo could have given: we don't feel the motion of the Earth around the Sun because the whole Earth is in free fall, so there is nothing to feel.

An important practical idea which we have seen in Galileo's work is that motion in different directions is *independent*: the vertical acceleration tells us the change in the speed in the vertical direction, and similarly for other directions.

In the next twelve chapters we will extend our exploration of gravity into the Solar System, to stars and to galaxies, but apart from Newton's law of gravitation we will not have to introduce any new ideas that do not already arise in terrestrial experiments. This is surely one of the most satisfying aspects of physics, that by drawing conclusions from experiments on the Earth we can make sense of what is happening in distant parts of the Universe. Not until Chapter 15, when we begin to get into special and general relativity, will we need some essentially new ideas.

Satellites: what goes up doesn't always come down

Many people assume that satellites orbit the Earth far above its surface, but the numbers tell a different story. Most satellites orbit at less than 300 km above the ground. Compared with the radius of the Earth, 6400 km, this is very small. Their orbits just skim the top of the atmosphere. We can expect, therefore, that the acceleration of gravity on such a satellite will not be very different from what it is near the ground. How then can it happen that the satellite doesn't fall to the ground like our cannonballs in the first chapter?

The answer is that it tries to, but the ground falls away as well. Imagine firing a cannon over a cliff. Eventually the ball will fall back to the height from which it was fired, but the ground is no longer there. The Earth has been cut away at the cliff, so the ball must fall further in order to reach the ground.

If we kept cutting the Earth away, the ball would just keep falling without ever hitting the ground. Now, the Earth is spherical, so it is already "cut away". Moreover, gravity attracts bodies toward the center of the Earth, so as the body moves around the Earth, the direction in which it is trying to fall keeps changing. Therefore, if we fire the cannonball fast enough, it might just keep falling toward the Earth without ever reaching the ground. It would then be a satellite.

It is not hard to get a rough idea of how fast a satellite has to be going in order to stay in orbit. In the first chapter, in Investigation 1.3 on page 5, we saw that the maximum range of a cannonball fired with speed V is V^2/g , where g is the acceleration of gravity, 9.8 m s^{-2} . If we set this range equal to the radius of the Earth, $R = 6400 \text{ km}$, and solve for V , then we must get a number which has about the right size. The result is that, as a first guess, V must be $(gR)^{1/2}$.

The calculation in Investigation 3.1 on page 22 shows that this guess is exactly right: the orbital speed of a satellite near the Earth's surface is

$$V_{\text{orbit}} = (gR)^{1/2}, \quad (3.1)$$

which is 7.9 km s^{-1} .

Given that the circumference of the satellite's orbit is 40 000 km, one orbit at this speed will take 5100 s, or 84 minutes. Since we have used values of g and R appropriate to the surface of the Earth, the true period of a typical near-Earth orbit at 300 km altitude is a bit longer, more like 91 minutes. (Geostationary orbits are much higher: see Chapter 4.)

We can solve Equation 3.1 for the acceleration g by squaring and dividing by R . If we change the symbol for acceleration to a (which will denote any acceleration – we reserve g to mean the Earth's surface acceleration) then we get

$$a = V_{\text{orbit}}^2/R. \quad (3.2)$$

This is the general expression for the acceleration of a particle that follows a circular orbit with radius R and speed V_{orbit} .

In this chapter: we use the equivalence principle to explain how satellites stay in orbit. We generalize the computer program of Chapter 1 to compute orbits of satellites.

► Communications and many weather satellites, which must be in "geostationary" orbits, are an important exception, being in distant orbits. We will return to these orbits in Chapter 4.

► The picture behind the words on this page is the Hubble Space Telescope (HST), a satellite launched by the National Aeronautics and Space Administration (NASA), with participation from the European Space Agency (ESA) as well. The HST has opened the most distant reaches of the Universe to view by making observations above the atmosphere of the Earth. Its low orbit makes it accessible to astronauts for repair and for upgrading its instruments. Image courtesy ESA.

In this section: we show how to do what Galileo did, namely to consider the motion and forces separately in vertical and horizontal directions. We distinguish the idea of speed from that of velocity.

Taking motion apart

Before looking at Investigation 3.1 on page 22, we will find it helpful to make a clear distinction between the words “speed” and “velocity”. I have used the two words virtually interchangeably up to now, but from now we should keep them separate. Let us illustrate the difference with an example, the cannonball of Chapter 1, which moves both vertically and horizontally.

First, let us be clear on what we mean by the vertical and horizontal motions. We could measure the cannonball’s horizontal motion, for example, by firing the cannon when the Sun is directly overhead, and then watching the ball’s shadow. By our discussion in Chapter 1, the shadow will move at a constant speed. Similarly, we could discover the vertical motion by shining a bright light at it from behind, and watching the cannonball’s shadow on a screen directly in front of it. This motion would be indistinguishable from that of any other ball fired with the same vertical speed, regardless of its horizontal speed.

Together, the two motions completely describe the body’s trajectory, and it is usual to give the *set* consisting of the horizontal speed and the vertical speed the name *velocity*. The word “speed” then refers to the rate of change of a distance with time, but the word “velocity” contains directional information: knowing the speed in both the vertical and horizontal directions, we can figure out the actual direction in which the cannonball is moving.

Mathematically, this set is called a **vector**. The vertical speed is called its vertical **component**. Thus, the components are just the usual numbers we have been considering. Mathematicians and physicists reserve the word “velocity” for the vector and “speed” for numbers associated with specific directions, and we shall do the same.

An important number associated with the velocity is the particle’s total speed, which means the number of meters per second the particle goes in its own direction, rather than its projection along one of the two directions. When I use the word *speed* on its own, rather than as horizontal speed or something like that, then I will mean total speed. By the **Pythagorean theorem**, the total speed is the square-root of the sum of the squares of the components of the velocity.

Acceleration, and how to change your weight

In this section: weightlessness in free fall, and weighing much more when a rocket takes off, are ways in which the equivalence principle changes the weight of astronauts.



Figure 3.1. An Airbus research aircraft about to go into free fall, which it can sustain for up to 22 s. Image courtesy Novespace/Airbus.

When astronauts orbit the Earth they are weightless: they float across their cabin, they release pencils or balls from their hands and watch them float too, they perform space-walks outside their spacecraft and do not require a tether in order to stay up there with the craft. It is tempting to conclude from their weightlessness that gravity is very weak so far from the Earth, but that can’t be right. We have seen that they are not particularly far away; and in any case, if gravity is so weak, why do they stay in orbit instead of flying off in a straight line? The real explanation is to be found in the equivalence principle.

The astronauts and their spacecraft are in free fall, so all the objects near them behave as if there were no gravity. Without gravity, things are weightless.

Being free of gravitational forces means having no weight. We shall see below that astronauts in orbit around the Earth are in free fall, constantly falling towards the Earth. Therefore, they are perfect examples of the hypothetical falling Galileo. When pens float alongside them, we see direct proof of the equivalence principle. A weightless environment can also be created for a limited time by allowing an airplane to fall freely. Research organizations use such aircraft regularly. (See Figure 3.1.)

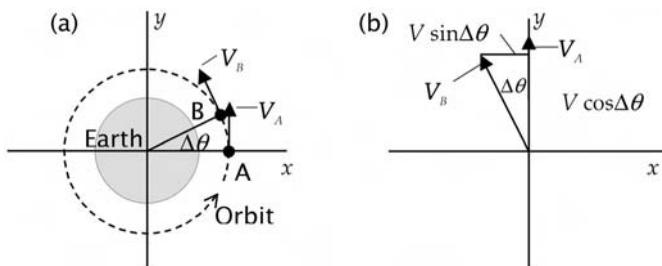


Figure 3.2. Velocities at two nearby points A and B on a circular orbit, as required for the calculation in Investigation 3.1 on the next page. The angle between the points is $\Delta\theta$. Velocities are drawn as arrows, whose length is equal to the total speed V . In panel (a) we see the circular orbit, the two points A and B, and the velocity arrow (tangent to the circle) at each point. The arrows have the same length but different directions.

In (b) we see the same velocity arrows, moved to the origin and magnified so they can be compared. The angle between them is the same as the orbital angle $\Delta\theta$ between points A and B. At point A the speed is only in the y-direction. At B it has components in both directions, which are marked in the diagram. Remember that both velocity arrows have length V .

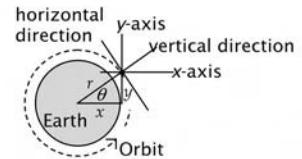
To fill out this explanation, let us ask what the sensation of weight really is. How do we *experience* our own weight? What we feel are the effects of the floor pushing up on us to support us against gravity. This force is transmitted to our bones and joints, which are somewhat compressed in our legs and extended in our arms as we stand. Our internal organs hang from their attachment points inside our chest and abdomen, and these supporting tissues are also stressed. All through our bodies there are nerve endings picking up these stresses and telling us that we are not weightless. If instead we find ourselves in free fall, these stresses disappear: no forces need be exerted by the supporting tissues to keep our lungs in place or our elbows together. The *sensation* of weight disappears completely.

The floor, not the force of gravity, is responsible for our weight.

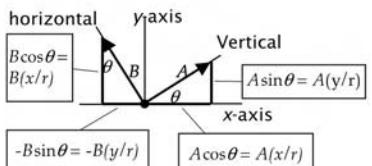
This argument can be turned around to explain why astronauts feel heavier when their rockets are firing, or indeed why we feel heavier when an ascending elevator starts up. Since gravity works only by inducing accelerations, any other steady acceleration mimics gravity. If a rocket accelerates at five times the acceleration of gravity (even in everyday language this is called “5 g’s”), its occupants will feel five times as heavy. Astronauts can be trained for this on a large centrifuge. Using Equation 3.3 on the next page for circular acceleration, we find that a centrifuge of 5 m radius creates an acceleration of 5g when its speed is 16 m s⁻¹, which corresponds to one revolution every two seconds. Anyone who has ridden a roller coaster or who has been unlucky enough to have been caught in strong turbulence in a high-flying airliner will have experienced the discomfort of such changes in weight first-hand, generated by accelerations that are only roughly 1g in size.

Notice that it is normally only possible to eliminate the effects of gravity in a small region. Two experimenters in free fall at different places on the Earth are falling on radial lines through the center of the Earth, so they are accelerating relative to one another. Indeed, when we begin to study Einstein’s point of view on gravity, we will see that the physical part of gravity is its non-uniformity.

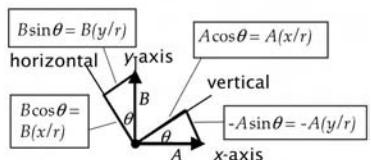
If gravity were everywhere uniform we could not distinguish it from acceleration. This is the sense of the word *equivalence* in *equivalence principle*. Therefore, the changes in gravity over distances tell us that we are really dealing with gravity and not simply a uniform acceleration. We will elaborate on this subject in Chapter 5.



(a) The relationship between the vertical and the x-axis depends on the angular position θ of the point on the orbit. The orbital radius r and the x- and y-coordinates shown here are used below.



(b) Taking a vertical vector (A) and a horizontal one (B) and finding their components in the x-y coordinates.



(c) Taking an x-directed vector (A) and a y-directed vector (B) and finding their horizontal and vertical components.

Figure 3.3. Coordinates for the satellite-orbit calculation in Investigation 3.2 on page 23. The diagrams show how to find the components of vectors from vertical/horizontal coordinates to the x-y system. The orientation of the vertical and horizontal coordinates is as in Figure 3.2. Expressions for components are given in boxes.

Investigation 3.1. Keeping a satellite in a circular orbit

Here we examine the geometry of a circular orbit in order to find the acceleration required to keep a satellite in uniform circular motion. Then we can set this equal to g to find the speed of a satellite orbiting near the Earth. In a later investigation we will test this result with a computer simulation of a satellite trying to get into orbit. But first we should look a little more at vectors, which were introduced in the text to describe velocity.

Other quantities may also be vectors. The position of a particle is given by a set of two numbers, which are its coordinates; this vector is called the **position vector**, or sometimes the **displacement** of the particle from the origin of coordinates. The acceleration is also a vector, but there are no special names for distinguishing between the acceleration vector and its components.

Now, suppose that the satellite's speed is V and its orbital radius is R . A small arc of the orbit is shown in panel (a) of Figure 3.2 on the preceding page. When the satellite is at point A its speed is in the direction shown by the arrow, whose length is V . When it gets to point B its speed hasn't changed, but the direction it is traveling in has, as shown. This change requires an acceleration, since its speed in both the x - and y -directions has changed.

In Figure 3.2 I show the geometry of the situation. Suppose for simplicity that the orbit lies in the plane of the Earth's equator. This is the plane of the diagram in the first panel of the figure. The x - and y -axes are rectangular coordinates in this plane, whose origin is at the center of the Earth. The orbit is a circle in this plane. The points A and B between which we wish to calculate the acceleration are indicated by dots on the orbit, and the velocity of the particle at each of these points is shown as an arrow. The length of the arrow is the same at the two points, but its direction has changed. The points are separated by an angle of $\Delta\theta$, which we eventually will assume is small. The velocity arrow has rotated by the same angle from A to B.

In the second panel, the velocities themselves are shown, magnified by a factor of two. What we want is to find the change in the speeds in the x - and y -directions. The velocity at A has an x -speed of zero and a y -speed of V . The parts of the velocity at B can be deduced by using the trigonometry of the right triangle whose hypotenuse is the velocity arrow of B, which has length V . Then the x -speed of this velocity is $-V \sin \Delta\theta$, and its y -speed is $V \cos \Delta\theta$. Thus, the change in the x -speed from A to B is $-V \sin \Delta\theta$, while the change in the y -speed is $V(\cos \Delta\theta - 1)$.

Exercise 3.1.1: Vectors

Quantities that have a value but no direction are called **scalars**. Decide whether the following physical quantities should be described mathematically by scalars or by vectors: (a) the mass of a rock; (b) the electric force on a charged particle; (c) the temperature of a room; (d) the slope of a hill.

Exercise 3.1.2: Period of a satellite orbiting near the Earth's surface

Use Equation 3.1 on page 19 to calculate the orbital period of a satellite near the Earth. Assume that the acceleration of gravity at the height of the satellite is the same as on the ground, $g = 9.8 \text{ m s}^{-2}$. Take the radius of the orbit to be the radius of the Earth, 6400 km, plus the height of the satellite above the Earth, 300 km.

From this figure it is also apparent that for very small $\Delta\theta$ the largest piece of the change of velocity is in the x -speed, which is directed perpendicular to the direction of motion at A, i.e. toward the center of the circle. This means that the *instantaneous acceleration of a circular orbit is directed towards the center of the orbit*. What we have to do is to figure out how big this acceleration is.

Since the satellite travels at (total) speed V , it moves once around the Earth in a time equal to the circumference of its orbit divided by its speed. This is the period of its orbit, P :

$$P = 2\pi R/V.$$

In a small time Δt , it will travel a fraction $\Delta t/P$ of the orbit, so the corresponding angle $\Delta\theta$ will be (in degrees)

$$\Delta\theta = 360^\circ \frac{\Delta t}{P}.$$

If this angle is very small, then it is clear from Figure 3.2 on the previous page that the change in the speed in the y -direction is very nearly equal to the arc of a circle of radius V subtended by the (small) angle $\Delta\theta$. This is simply the circumference of such a circle times the fraction ($\Delta\theta/360^\circ$). The result is

$$\begin{aligned} \text{change in } y\text{-speed} &= 2\pi V \frac{\Delta\theta}{360^\circ} = 2\pi V \frac{\Delta t}{P} \\ &= 2\pi V \frac{\Delta t}{2\pi R/V} = \frac{V^2}{R} \Delta t. \end{aligned}$$

The acceleration is this change divided by Δt ,

$$a = \frac{V^2}{R}. \quad (3.3)$$

This is the desired formula. Notice that it does not contain Δt or $\Delta\theta$. This is because we made various approximations that were valid only if these were sufficiently small. This means that Equation 3.3 is the *exact expression for the instantaneous acceleration*.

For a satellite orbiting near the Earth, this acceleration will be just g . Putting this into Equation 3.3 and solving for V gives Equation 3.1 on page 19.

Getting into orbit

In this section: how satellites get into and stay in orbit.

I remarked in Chapter 2 that the orbits of planets are ellipses. This property also applies to satellites of the Earth, the circular orbit being a special case of an ellipse. An important feature is that the orbit is *closed*: a satellite will always return to the place where it was set into orbit. More than that, when it returns to that spot it will be moving in the same direction as before. Therefore if we were to try to launch a cannonball into orbit just by firing it at a sufficiently high speed, we would not succeed: its elliptical orbit would make it want to return to the launching point from *below*. It would necessarily hit the ground somewhere else as it tried to pursue its orbit into the interior of the Earth.

To get something into orbit, one has to give it at least two pushes: one to get it off the ground, and a second to put it into an orbit that does

Investigation 3.2. Achieving orbit

Here we ask the computer to calculate for us the orbit of a satellite that is launched from a height of 300 m (say from the top of a cliff), and given a perfectly horizontal velocity. Does the satellite hit the Earth, or does the Earth fall away faster than the orbit does?

The calculation is basically the same as the trajectory program in Chapter 1, with the one difference that the acceleration of the body depends on where it is. As we saw in Investigation 3.1, the acceleration must always point towards the center of the Earth. We will make it slightly easier to calculate the orbit by assuming that the total magnitude of the acceleration is always g , the same as the acceleration on the surface of the Earth. This is not quite right, since gravity gets weaker as we get further from the surface, but we will not take that into account until we do the orbit program for the Solar System in Chapter 4. By taking the acceleration to be constant, we will not get the right orbits for a particle launched faster than the circular velocity, since that particle will move further away from the Earth.

Let us call a_x and a_y the accelerations in the x - and y -directions, respectively, at any time t . (In the program EarthOrbit they are called ax and ay .) From Figure 3.3 on page 21, we can see that, at a point of the orbit given by the angle θ , a line of total length g that points towards the center of the circle has x -length $g \cos \theta$. Since the cosine of this angle is, by the first part of the figure, $\cos \theta = x/r$, this is gx/r . But the acceleration is directed *towards* the center, so it has to be given a negative value: the change of x -speed produced by this acceleration is negative because it is directed in the negative- x sense. Putting all this together gives

$$a_x = -gx/r.$$

Similarly, for the acceleration in the y -direction, the geometry shown in Figure 3.3 on page 21 shows that

$$a_y = -gy/r.$$

We take $g = 9.8 \text{ m s}^{-2}$.

As we did in Chapter 1, we will find approximate changes in the x - and y -speeds during a small interval of time Δt by multiplying the accelerations in these directions at time t by Δt . If we let the x -speed be v (v in the program) and the y -speed be u (u in the program), then the approximate change in v is

$$\Delta v = a_x(t)\Delta t, \quad \text{and} \quad \Delta u = a_y(t)\Delta t. \quad (3.4)$$

In the program, we denote Δv by `deltaV`, and similarly for other variables. We also need the calculate the changes in positions, which are determined by the speeds. We have

$$\Delta x = v(t)\Delta t, \quad \text{and} \quad \Delta y = u(t)\Delta t. \quad (3.5)$$

Recall that in Investigation 1.3 on page 5, we saw that this method gave the exact result for a cannonball trajectory. Unfortunately, we cannot expect this to be true here, as well, because in our case the acceleration changes from point to point. Therefore, as we have discussed before, we have to take a small enough time-step to give an accurate result. The program on the website uses a time-step of 0.4 s, and gets a fairly good circular orbit when we take the initial speed to be exactly the orbital speed given by Equation 3.1 on page 19. By changing the time-step you can explore the question of how accurate the calculation is.

Another important feature of this program (and an essential element of any computer program!) is how it stops. We have given it a way of knowing when the orbit has gone once around, so it stops if the orbit hits the Earth or after it goes around once. The logic is explained on the website.

An important point about this calculation is that we perform all our calculations with the x -speed and the y -speed, not the vertical and horizontal parts of the velocity. The reason is Galileo's principle of relativity, which was incorporated into Newton's first law, as we discussed in Chapter 2. The independent motions of a body are those that would continue unchanged in the absence of external forces, namely those along straight lines. Since the vertical and horizontal directions change with time, they cannot be treated separately. But the x - and y -motions are separate, and they depend only on the accelerations in those directions. All our later computer programs for the motion of bodies will consistently use these rectangular coordinates.

The conversion of x - y speeds and positions to vertical-horizontal ones depends on where in the orbit one is. From Figure 3.3 on page 21 one can deduce the following conversions for the components of any of the vector quantities (distance, speed, acceleration):

$$\begin{aligned} (x) &= (\text{vert}) \cos \theta - (\text{horiz}) \sin \theta \\ (y) &= (\text{vert}) \sin \theta + (\text{horiz}) \cos \theta \\ (\text{vert}) &= (x) \cos \theta + (y) \sin \theta \\ (\text{horiz}) &= -(x) \sin \theta + (y) \cos \theta. \end{aligned} \quad (3.6)$$

Here the notation is that (x) stands for the x -component of the quantity involved, (vert) for its vertical component, and so on. As the boxes in Figure 3.3 on page 21 show, the trigonometric functions in the above equations may be replaced by expressions using the coordinates of the point on the orbit:

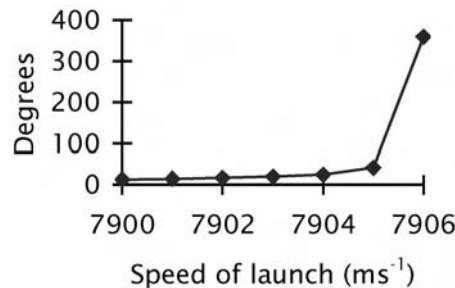
$$\sin \theta = \frac{y}{r}, \quad \text{and} \quad \cos \theta = \frac{x}{r}.$$

not collide with the Earth. That is why the “boost phase” of a satellite launch is always followed by a crucial “orbit insertion” event in which rockets are fired again. We cannot launch satellites just by shooting them out of a cannon.

Most satellite launches use more phases of acceleration than just the boost and insertion. Usually the boost is divided into two or three phases, contributed by different stages of the rocket. This is not required by the properties of orbits in a gravitational field. Rather, it is useful for keeping the amount of fuel used to a minimum. Once a certain amount of fuel has been used, the empty fuel tanks are a useless weight for the rocket, so it is more efficient to drop them off and use the remaining fuel to propel less weight into orbit.

In order to illustrate the expression for the circular orbital speed that we derived above, and to develop the computer techniques and programs that we will need in later chapters to look at orbits around the Sun, binary and multiple star systems, and orbits around black holes, we discuss a simple satellite-launching problem in Investigation 3.2. We imagine that the satellite is launched from a certain height in

Figure 3.4. A few attempts at getting into orbit, as calculated by the program EarthOrbit. The “satellite” is fired horizontally from a height of 300 m. The trajectory stops where it hits the Earth, which is taken to be a perfect sphere of radius 6 378 200 m. We do not show a picture of the trajectory, since it is so close to the Earth that different trajectories would all come out superimposed. Instead, we plot the angle through which the orbit turns before the trajectory hits the Earth, as a function of the launch speed. Note how much difference small changes in this speed make. The final speed, 7906 m s^{-1} , is the one that attains orbit.



a horizontal direction. With too little speed, it falls and hits the Earth. The angular distance around the Earth that the satellite travels for a given launch speed is shown in Figure 3.4. Reaching orbit is remarkably sensitive to the exact speed. With a launch speed of 7900 m s^{-1} the satellite hardly gets anywhere. If its launch speed rises to only 7906 m s^{-1} , the satellite goes into orbit.

The Solar System: a triumph for Newtonian gravity

As children of our age, we find it natural to think of the planets as cousins of the Earth: remote and taciturn, perhaps, but cousins nevertheless. To visit them is not a trip lightly undertaken, but we and our robots have done it. Men have walked on the Moon; live television pictures from Mars, Jupiter, Saturn, Uranus, and Neptune have graced millions of television screens around the world; and we know now that there are no little green men on Mars (although little green bacteria are not completely ruled out).

Among all the exotic discoveries have been some very familiar sights: ice, dust storms, weather, lightning, erosion, rift valleys, even volcanos. Against this background, it may be hard for us to understand how special and mysterious the planets were to the ancients. Looking like bright stars, but moving against the background of "fixed" stars, they inspired awe and worship. The Greeks and Romans associated gods with them, and they played nearly as large a part in astrology as did the Moon.

It was a giant step forward for ancient astronomers, culminating in the great Greek astronomer Ptolemy, to show that what was then known about planetary motions could be described by a set of circular motions superimposed on one another. These were called **cycles** and **epicycles**. It was an even greater leap for Copernicus to argue that everything looks simpler if the main circular orbits go around the Sun instead of the Earth. Not only was this simpler, but it was also a revelation. If the Earth and the planets all circle the Sun, and if the Earth is simply in the third orbit out, then probably the planets are not stars at all, and the Earth might be a planet, too.

This probability became a virtual certainty when Galileo trained his first telescope on the night sky. Not only did he discover that Jupiter had moons, just like the Earth, only more of them, he also saw craters and mountains on the Moon, indicating that it was rocky like the Earth; and he saw the phases of Venus, the shadow that creeps across Venus as it does the Moon as these bodies change their position relative to the Sun. This meant that Venus was a body whose size could be measured: it was not a mere point-like star. When Kepler (whom we met in Chapter 2) showed, by extraordinarily painstaking calculations, that the planetary orbits were actually ellipses, the modern *description* of the motion of the planets was essentially complete.

But Galileo's study of motion on Earth soon raised an even bigger problem. Since bodies travel in straight lines as a rule, and since the planets do not, what agency forces them to stay in their orbits? Newton saw this problem clearly, and he had the courage to say that it needed no extra-terrestrial solution: gravity, the same gravity that makes apples fall from trees, also makes the planets fall toward the Sun. But what was the *law* of gravity? What rule enables one to calculate the acceleration of the planets?

In this chapter: applied to the Solar System, Newton's new theory of gravity explained all the available data, and continued to do so for 200 years. What is more, early physicists understood that the theory made two curious but apparently unobservable predictions: that some stars could be so compact that light could not escape from them, and that light would change direction on passing near the Sun. Einstein returned the attention of astronomers to these ideas, and now both black holes and gravitational lenses are commonplace.

▷This name is pronounced "Tolomey". His full name was Claudio Ptolomaeus, and he lived in Alexandria during the second century AD. Little else is known of him.

▷The image behind the text on this page is from a beautiful photograph of the Moon taken by the Portuguese amateur astronomer A Cidadao on 1 March, 1999. Used with permission.

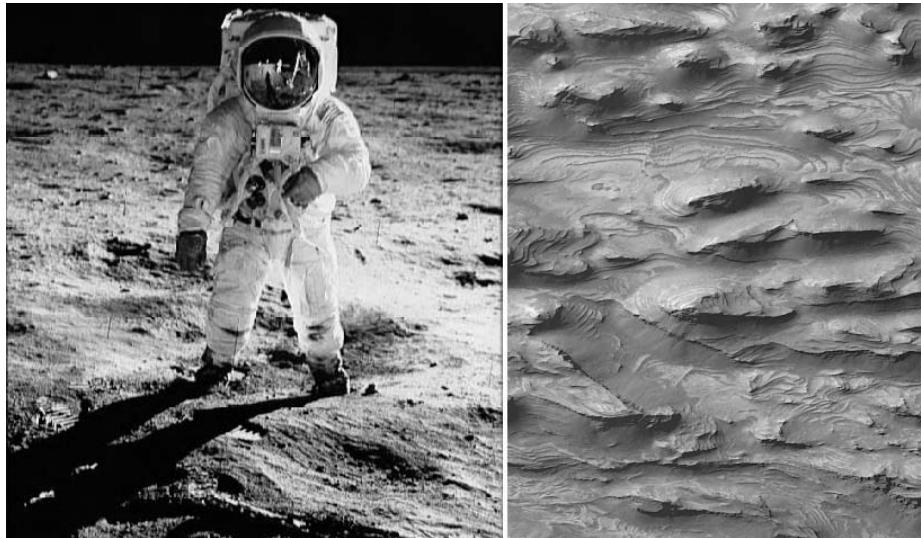


Figure 4.1. Both the Moon (left) and Mars (right) appear to be desolate, uninhabited deserts. But Mars appears to have experienced erosion, probably by flowing water, at some time in its past. Left image courtesy NASA; right courtesy NASA, the Jet Propulsion Laboratory (JPL), and Malin Space Science Systems.

In this section: we learn how knowledge of the Moon's distance, which was available to Newton, makes the law of gravity that he invented very plausible.

How to invent Newton's law for the acceleration of gravity

Let us look first at the nearest “planet”: the Moon. If the laws of mechanics postulated by Newton are to apply in the heavens as well, then we should be able to deduce from the motion of the Moon what the force on it is. Suppose that the Moon is in a circular orbit. This is a good first approximation, but we shall have to return to the question of elliptical orbits later. We have seen in Chapter 3 that, for circular motion at speed V and radius R , the acceleration a is V^2/R . We can eliminate V in terms of the radius of the orbit R and the period P , because the speed is just the distance traveled (circumference of the orbit) $2\pi R$ divided by the time taken, P . This means that the acceleration is $(2\pi/P)^2 R$ towards the Earth.

Now, Newton knew the distance R to the Moon; even the Greeks had a value for it, by measuring its **parallax** from different points on the Earth, as shown in Figure 4.2. Since Newton also knew the period P of the Moon's orbit, he could work out its acceleration.

If we consult Table 4.1 for the modern values of these numbers, we can calculate from Equation 3.1 on page 19 or Equation 3.3 on page 22 that the acceleration of the

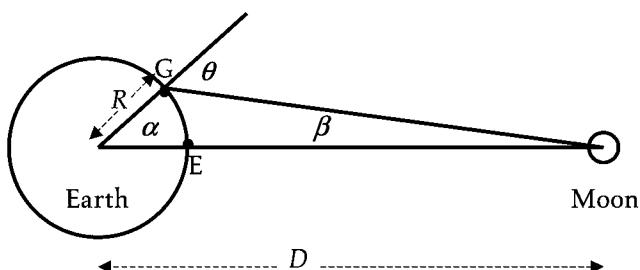


Figure 4.2. The direction to the Moon is different at different places on the Earth; this is called the Moon's parallax. One can use the parallax to determine the Moon's distance D . Suppose the Moon is directly overhead at a point E on the Earth, and suppose one measures its position in the sky from point G at the same moment, obtaining that it lies at an angle θ down from the vertical there. The point G is known to be more northerly on the Earth than E , by a latitude angle α . Simple geometry then says that the angle between E and G as seen from the Moon is $\beta = \theta - \alpha$, and the angle of the triangle opposite the desired distance D is $180^\circ - \theta$. By the law of sines for triangles, we have $\sin(180^\circ - \theta)/D = \sin \beta/R$, where R is the radius of the Earth. This can be solved for D . Ptolemy performed essentially this measurement of θ , and knew α and R reasonably accurately. He deduced that the Moon's distance was 59 times the Earth's radius. The right value is just over 60.

Table 4.1. Data for the Moon.

Average distance from the Earth, R (km)	Period of orbit, P (s)	P^2/R^3 ($s^2 \text{ km}^{-3}$)	Average speed (km s^{-1})	Mass (kg)
3.84×10^5	2.36×10^6	9.84×10^{-5}	1.02	7.3×10^{22}

Moon is 0.0027 m s^{-2} . Newton reasoned that if this was due to the Earth's gravity, then (like any gravitational acceleration) it could not depend on the Moon's mass, so it could depend only on how far the Moon was from the Earth. In particular, he guessed that it might depend only on how far it was from the *center* of the Earth. Compared with the acceleration of gravity on the Earth's surface, 9.8 m s^{-2} , that of the Moon is smaller. How much smaller? The ratio of the two accelerations, $9.8/0.0027$, is 3600. The ratio of the radius R of the Moon's orbit, 384 000 km, to the radius of the Earth itself, 6380 km, is 60.

Clearly, the ratio of the accelerations is the *square* of the ratio of the distances taken in the opposite sense: the acceleration produced by the Earth is inversely proportional to the square of the distance from the Earth's center.

We have already seen in Equation 2.3 on page 13 that the simplest form of such an inverse-square law of gravity that obeys the equivalence principle is

$$F_{\text{grav}} = \frac{GM_1M_2}{r^2}, \quad (4.1)$$

where M_1 in this case is the mass of the Earth and M_2 that of the Moon.

One can imagine how Newton might have reacted to this result: such a simple relation cannot be coincidence! Surely it must also apply to the planets in their orbits around the Sun. But Newton knew that the orbits of the planets were ellipses, not simple circles. Could that also be a consequence of the inverse-square law? That is what we turn to next.

The orbits of the planets described by Newton's law of gravity
 In Table 4.2 on the following page I have listed the main properties of the planets and their orbits. Here we encounter for the first time the astronomer's unit of distance in the Solar System, the **astronomical unit**, denoted AU. It is defined as the average distance of the Earth from the Sun, $1.496 \times 10^{10} \text{ m}$. In these units, distances become easier to comprehend: 2 AU is just twice the radius of the Earth's orbit, but what is $3 \times 10^{10} \text{ m}$? (It is just 2 AU again.)

Now look at column 4, where I have calculated the ratio of the square of the period to the cube of the average distance from the Sun. Kepler had noticed that these values were remarkably similar for all the known planets (out to Saturn), and we now call this *Kepler's third law*. (He didn't know the absolute distances between planets very well, but could deduce their ratios from observations, and that was enough to deduce the constancy of this number.)

Newton recognized that this strange relation provides the crucial evidence that the gravitational force does indeed fall off as the square of the distance. Again we consider an idealized circular orbit, for simplicity. From Equation 3.2 on page 19, the quantity in column 4 is, in terms of the acceleration a and the radius R ,

$$\frac{P^2}{R^3} = \frac{4\pi^2}{R^2 a}. \quad (4.2)$$

If this is to be constant, then $R^2 a$ must be constant, or a must be proportional to $1/R^2$, exactly as we inferred from the Moon's orbit.

In this section: given Newton's law of gravity and the distances to the planets, the law must predict all the details of their orbits. This is the critical test that the law had to pass before Newton would believe it.

Notice that in this calculation we had no “adjustable” parameters: if the data hadn’t fit our proposed law we would have had to throw the law away. But they *did* fit: surely we are on the right track to an understanding of the planetary orbits.

Encouraging as this argument is, it does rely on the idealization of circular motion. Moreover, the numbers in column 4 are not perfectly constant: there are small but measurable deviations. Before he could convince himself of his law of gravity, Newton needed to show that a $1/R^2$ acceleration also produced the elliptical orbits that Kepler had observed, and that the deviations in column 4 could be predicted from the ellipticity of the orbit, something much more complicated. To do this Newton had to invent the calculus.

With a personal computer we can come close to providing a demonstration of this result in a matter of minutes, without calculus. The website contains a program which calculates planetary orbits. It is based on the trajectory program of Chapter 1, adapted to the planetary case by the calculations in Investigation 4.1. It produced the orbit of Mercury shown in Figure 4.3.

Our computed orbit closes smoothly on itself, and it “looks” elliptical. The **eccentricity** of the ellipse may be calculated by estimating the ratio, q , of the maximum to the minimum distance from the Sun:

$$e = \frac{1 - q}{1 + q}. \quad (4.3)$$

Remember that the Sun sits at a focus of the ellipse, not its center. So the ratio q is not the same as the ratio s of the short axis to the long axis. $e = (1 - s^2)^{1/2}$.

My rather crude estimate of q from the graph gives $e = 0.21$, which is acceptably close to the observed eccentricity, 0.206 (i.e. the ratio of Mercury’s short axis to its long one is 0.98). Apart from Pluto, Mercury has the most elliptical orbit of all the planets.

I started the computer calculation of Mercury’s orbit at its minimum distance from the Sun (4.6×10^7 km), which is called its *perihelion* distance. To compute the orbits of other planets you may use the data in Tables 4.2 and 4.3. A nice thing to try is the orbits of Neptune and Pluto, because, as the tables make clear, they cross!

Of course, a numerical calculation cannot *prove* that the orbits are perfect ellipses, because there is always some inaccuracy in the fact that we take finite time-steps; instead of changing smoothly, the acceleration changes in discrete steps. This calculation cannot be a substitute for Newton’s proof using calculus. Nevertheless,

Table 4.2. Planetary data. One year is 3.1557×10^7 s. The number of known satellites of most of the outer planets is likely to go up with further exploration. For comparison, the mass of the Sun is 1.989×10^{30} kg, and the radius of the Sun is 6.9599×10^5 km. Astronomers are discovering an increasing number of small planetary bodies outside the orbit of Pluto, but none as large as Pluto. They are not usually called planets, and are not in this table.

	1 Average distance from the Sun, R (10^6 km)	2 Period, P (AU)	3 P^2/R^3 (10^{-10} s 2 km $^{-3}$)	4 Average orbital speed (km s $^{-1}$)	5 Mass (10^{24} kg)	6 Known satellites
Mercury	57.9	0.387	0.241	2.98	47.9	0.33
Venus	108.2	0.72	0.615	2.97	35.1	4.87
Earth	149.6	1.00	1.00	2.97	29.8	5.97
Mars	228.0	1.52	1.88	2.97	24.1	0.642
Jupiter	778.3	5.20	11.86	2.97	13.1	1900
Saturn	1429.4	9.54	29.46	2.99	9.65	568.41
Uranus	2871.0	19.22	84.01	2.98	6.80	86.83
Neptune	4504	30.06	164.1	2.97	5.43	102.47
Pluto	5913.5	39.5	247.0	2.97	4.74	0.0127
Moon	0.384	(from Earth)	0.0748	9.84×10^5	1.02	0.073

Investigation 4.1. How to follow the orbit of a planet

The calculation of the orbit of a planet is very similar to the one for the satellite around the Earth. The only essential change is in the law for the acceleration. Here we adopt the law that the acceleration is towards the Sun (which we place at the origin of coordinates) and has magnitude

$$a = k/r^2. \quad (4.4)$$

The distance to the Sun is given by r . The constant k can be inferred from Kepler's third law, the constancy of the values in column 4 of Table 4.2. The way to do this is suggested in Exercise 4.1.1.

Newton was able to estimate k for the Earth-Moon system rather crudely from the then available estimate of the distance to the Moon.

The x - and y -accelerations are similar to those used in Investigation 3.2 on page 23, but with g replaced by k/r^2 :

$$a_x = -k \frac{x}{r^3}, \quad (4.5)$$

$$a_y = -k \frac{y}{r^3}. \quad (4.6)$$

By changing the program EarthOrbit to use this as the acceleration, we could obtain a working program for planetary orbits. Giving it initial data from Table 4.3 on page 32 produces good orbits, if the time-step is small enough. Exercise 4.1.3 suggests that you do this and compare the periods and eccentricities of the orbits you compute with the data in table.

In Exercise 4.1.4 you will see that the program is not terribly accurate unless you take small time-steps. There are two important further changes we will make to turn EarthOrbit into the more accurate Orbit. For readers who are interested in how good computer programs are constructed, Investigation 4.2 contains a description of these refinements: a predictor-corrector to make each time-step of the calculation more accurate, and a time-step halver to maintain the accuracy of the finite-difference method at each time-step as the orbital conditions change. These changes are not strictly necessary for planetary orbits, but they will allow us to use the program later for the more demanding problems of binary stars and multiple stars. The resulting program also does very well for extreme Solar System orbits, such as those followed by comets.

Exercise 4.1.1: Inverse-square-law constant

From Equation 4.2 on page 27, show that the constant k in Equation 4.4 is $(2\pi)^2 / (P^2/R^3)$. Evaluate this to give $1.327 \times 10^{11} \text{ km}^3 \text{ s}^{-2}$.

Exercise 4.1.2: Measuring the mass of the Sun and the Earth

The Newtonian law of gravity, Equation 4.1 on page 27, tells us the force on a body of mass M_2 exerted by the Sun (mass M_1 in the equation). Combine this with Newton's second law, $F = ma$, to show that the acceleration of the body of mass M_2 is $a = GM_1/r^2$. Use this to show that the force-law constant k in Equation 4.4 is $k = GM_1$. Convert this value of k to more conventional units using meters to find $k = 1.327 \times 10^{20} \text{ m}^3 \text{ s}^{-2}$. (Hint: since $1 \text{ km} = 10^3 \text{ m}$, it follows that $1 = 10^3 \text{ m km}^{-1}$. Multiply by the cube of this form of the number 1 to convert the units for k .) Now use the value of $G = 6.6725 \times 10^{-11} \text{ m}^3 \text{ s}^{-2} \text{ kg}^{-1}$ to find the mass of the Sun. Do a similar calculation for the value of Kepler's constant for the Moon, given in Table 4.2, to find the mass of the Earth.

Exercise 4.1.3: Simple orbit simulations

Run the orbit program as modified in this investigation for the planets Mercury, Jupiter, Neptune and Pluto. Start with the perihelion distance and speed given in Table 4.3 on page 32 and infer the value of k for each planet using Equation 4.2 on page 27 and the data in Table 4.2. Compute at least one full orbit. You will have to choose a time-step that allows the planet to move only a small distance at each step, and a number of time-steps that allow the planet to go all the way around. Use Equation 4.3 to calculate the eccentricity of the orbit. See how close you come to the eccentricity given in Table 4.3. Compare your orbit for Mercury with that shown in Figure 4.3.

Exercise 4.1.4: Assessing the accuracy of the simple orbit program

For the planet Mercury compute ten successive orbits and see if they lie on top of each other. They should do this, so the extent to which they do not reflects the inaccuracy of the approximations in the computer program. Reduce the time-step size, increasing the number of time-steps accordingly. Do the successive orbits lie more accurately on top of one another? As an estimate of the error, estimate the angle that the perihelion position of the orbit has moved after ten orbits. Calculate this error for different time-steps. Plot the error against the time-step. Is there a simple relation between them?

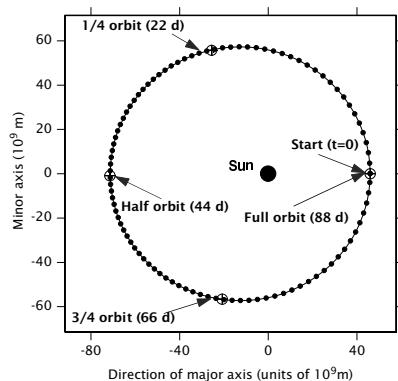


Figure 4.3. Simulation of the orbit of Mercury using the computer program Orbit from the website.

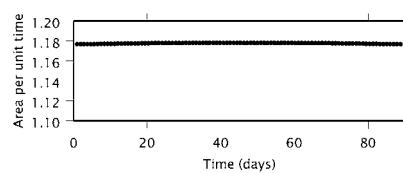


Figure 4.4. Kepler's area law computed for the orbit of Mercury. Each dot is the area of the triangle between the Sun and two successive dots on Mercury's orbit, divided by the time it takes Mercury to move between them, in units of $10^{20} \text{ m}^2 \text{ per day}$.

Investigation 4.2. A more sophisticated and accurate orbit program

This investigation is for readers who want to learn some of the ingenious methods by which computer experts improve the accuracy of computer programs. We will explore two improvements. One is to improve the accuracy of each time-step, and the other is to control the accuracy of each time-step.

(a) *Improving accuracy.* Recall that in Chapter 1, we found that the most sensible approximation to the change in the position of a body during a small interval of time Δt is obtained by *averaging* the speeds at t and $t + \Delta t$ and multiplying by Δt (Equation 1.5 on page 4). By the same reasoning, it would be sensible to find the change in a body's velocity by averaging its acceleration. We did not need to do this in Chapter 1, because there the acceleration g was constant. We did not do it in Chapter 3, even though for the program EarthOrbit the acceleration did not have a constant direction. There we just took the change in velocity to be the acceleration at time t times Δt (Equation 3.4 on page 23). This kept the program simple, but not as accurate as it could be. So here we would like to use

$$\Delta v = \frac{1}{2}[\alpha_x(t) + \alpha_x(t + \Delta t)]\Delta t, \quad (4.7)$$

and

$$\Delta u = \frac{1}{2}[\alpha_y(t) + \alpha_y(t + \Delta t)]\Delta t. \quad (4.8)$$

Similarly, the best way of computing the changes in the position of the body is to find the average of the speeds over the interval:

$$\Delta x = \frac{1}{2}[v(t) + v(t + \Delta t)]\Delta t, \quad (4.9)$$

and

$$\Delta y = \frac{1}{2}[u(t) + u(t + \Delta t)]\Delta t. \quad (4.10)$$

Although we would *like* to use these equations, there is a difficulty that we did not have to face in Chapter 1: here the acceleration depends upon the position, so we cannot calculate, say, $\alpha(t + \Delta t)$ without knowing $x(t + \Delta t)$, but we cannot calculate $x(t + \Delta t)$ without knowing $v(t + \Delta t)$, for which we need the acceleration $\alpha_x(t + \Delta t)$. Are we trapped in a circle with no exit? Not if we remember that with a computer we are only solving the equations with a certain accuracy, not exactly.

The method we will use is called a predictor–corrector technique. The idea is to guess, or “predict”, the values of, say, x and y at the later time by just multiplying the speeds at time t by Δt ; these are the positions that we used in EarthOrbit. Then we use this (admittedly somewhat incorrect) position to calculate α_x and α_y at $t + \Delta t$. Although they are not exactly right, they should be better than not correcting for the change of the acceleration with position. These values can then be used in Equation 4.7–Equation 4.8 to find v and u at $t + \Delta t$. Now comes the beautiful step: these can in turn be used in Equation 4.9–Equation 4.10 to find *new* values for x and y at $t + \Delta t$. These are the “corrections” to the first “predictions”. Since the corrected positions are calculated from better values of the acceleration and velocity, they should be better than the predicted positions.

Exercise 4.2.1: Assessing the accuracy of the improved method

Repeat Exercise 4.1.4 on the preceding page with the improved program Orbit. In particular, does the error depend in a different manner on the *average number* of time-steps?

We need not stop here. We can use the corrected positions as *new* predictions to give better accelerations, thence better velocities, and thence even better corrected positions. For person using a hand-calculator, predictor–corrector is tedious and time-consuming. But a computer is good at doing things repetitively. It is easy to program the computer to repeat this procedure as often as we wish. We just have to tell the computer when to stop making new corrections. We tell the computer to compare the prediction and correction at each stage, and when their difference is smaller than some predetermined accuracy level that we have given to the computer, then the process stops. This insures that Equation 4.7–Equation 4.10 are satisfied to whatever accuracy we desire. The program Orbit includes this feature. Mathematicians call this technique “iteration”. If successive changes in the predicted position become smaller and smaller, we say the method “converges” to the right answer. Note that it only converges to a solution of Equation 4.7–Equation 4.10. We are still left with the possibility that the time-step we chose in these equations was too large for these equations to give a good approximation to the real orbit. Fixing this problem is the aim of the next improvement.

(b) *Uniform accuracy.* The second new feature is the adjustment of the time-step to maintain a uniform accuracy. Consider what happens on a very eccentric orbit, such as that of a comet. Its speed is slow when it is far from the Sun, but it begins to move much more rapidly as it gets closer in. As we saw in Investigation 1.3 on page 5, we can only trust our finite-difference methods for computer programs if important physical quantities do not change much during the time-step. If we have a fixed time-step Δt , we expect much greater accuracy in our calculation of the orbit far away than near the Sun, where the position changes much more during Δt . Similar remarks apply to the accuracy of the velocity calculation if the acceleration changes by a large amount: if this happens, then even Equation 4.7 will not give accurate results.

Since the predictor–corrector method already looks ahead at the predicted position at time $t + \Delta t$ of the planet, we can look at the acceleration there to see if it is very different from the value at time t . If the difference between the accelerations at the original and the predicted positions is more than some preset fraction of the original value of the acceleration, then we should take a smaller time-step. I have written the program so that it goes back to the time t and cuts the time-step Δt in half. The test for the change in the acceleration is then applied again. The halving goes on and on until a satisfactorily small time-step is reached. This ought to give uniform accuracy over the whole orbit.

However, the way I have implemented this idea is crude, because at the next time-step its size reverts to the original one given as part of the data for the calculation. Since halving the time-step takes computing time, the program may run slowly for a highly eccentric orbit. You might like to see if you could improve the halving routine to make the program Orbit run faster.

by taking a small time-step and setting the accuracy parameters in the computer program to be small, we can make ourselves very confident of the result.

What is the value of G ?

Newton's law of gravity contains the constant of proportionality that we now call G , Newton's constant of gravitation. In fact, Newton did not actually know the value of G . Remember that the acceleration of the Moon is the force on it divided by its mass. If the Moon has mass m and the Earth mass M , then the force between them will be GmM/r^2 , so the Moon's acceleration will be GM/r^2 . Newton could deduce the product GM from observations of gravitational accelerations. (We can see how to do this in detail in Investigation 4.1 on page 29.) He could only deduce G if he could independently estimate M , the Earth's mass.

Newton actually tried to do this by assuming that the Earth's density was five times that of water, which he felt was a reasonable guess based on the density of rocks on the Earth's surface. Multiplying this density by the Earth's volume gave him a value for G of about $10^{-10} \text{ m}^3 \text{ s}^{-2} \text{ kg}^{-1}$. This is not bad, considering the data available to him. The value of Newton's gravitational constant is now measured to be 0.667×10^{-10} in these units. We can turn his argument around and use it to deduce the mass of the Earth from our values for g and the modern value of G . If M is the Earth's mass and R its radius, then Newton's law of gravity says that the downward acceleration of an object at the surface of the Earth is

$$g = \frac{GM}{R^2},$$

which means that

$$M = \frac{gR^2}{G}.$$

Using $g = 9.8 \text{ m s}^{-2}$ and $R = 6400 \text{ km}$, we find $M = 6.0 \times 10^{24} \text{ kg}$. (We use essentially the same method to deduce the Earth's mass from the Moon's acceleration in Exercise 4.1.2 on page 29.)

This is all very well if we know G , but how is G measured? The only way to separate G from M in the gravitational acceleration is to measure the gravitational acceleration produced by a body of known mass. Early attempts at this used mountains as the "known" mass: the direction that a plumb bob hangs will be slightly affected by the gravitational pull of a nearby mountain, and this is measurable by comparing the directions of plumb bobs on either side of the mountain. This isn't very accurate, however, and the modern method is due to the Englishman Henry Cavendish (1731–1810) (later Lord Cavendish), who succeeded in 1798 in measuring the mutual gravitational attraction of two balls in the laboratory. The force is very small, and his experiment was a marvel of precision physics for its day.

Even today, the measurements of G are accurate to only slightly better than one part in a thousand. By contrast, the product GM for the Sun is known to one part in one hundred million, from accurate tracking of interplanetary space probes' orbits. So the limit on the accuracy with which we know the mass of the Sun or any of the planets is the inaccuracy of G .

The reason for the relative imprecision of measurements of G directly is the weakness of the gravitational force between laboratory-sized objects, compounded by the difficulty of knowing exactly what the mass M of the laboratory mass is, and how it is distributed within the body. Real metal balls, for example, are never really

In this section: Newton introduced the proportionality constant G , but astronomical measurements could not determine its value. Only Earth-based experiments with objects of known mass could tell Newton the value of G .

Table 4.3. More data on the planets. The eccentricity is defined in terms of the ratio s of the minor and major axes of the ellipse by $e = (1 - s^2)^{1/2}$. The **perihelion** distance is the closest a planet gets to the Sun. Its maximum speed occurs at perihelion, where it is traveling perpendicularly to the direction to the Sun. Note that at perihelion, Pluto is closer to the Sun than Neptune, but its greater speed there carries it on a more eccentric orbit that is mostly outside Neptune's orbit. Pluto's peculiar orbit may have something to do with Neptune. Notice that its orbital period is exactly 3/2 that of Neptune.

	Eccentricity of the orbit e	Perihelion distance (10^6 km)	Maximum speed (km s^{-1})
Mercury	0.206	46.0	59.22
Venus	0.007	107.5	35.34
Earth	0.017	147.1	30.27
Mars	0.093	206.8	26.22
Jupiter	0.049	740.3	13.52
Saturn	0.053	1349.0	10.15
Uranus	0.046	2735.0	7.105
Neptune	0.012	4432.0	5.506
Pluto	0.249	4423.0	6.17

uniform and **homogeneous** inside, and as we will see later in this chapter, any non-sphericity will affect the force they exert.

Kepler's laws

In this section: Kepler's laws for planetary motion follow from Newton's. We can prove that to ourselves using computer programs to follow the orbits of planets.

We have come across Kepler's third law. What about his first and second?

Kepler's second law is the observation that planets follow orbits that are ellipses with the Sun at one focus.

We have already seen in Figure 4.3 on page 29 that Mercury's orbit is an ellipse, so it is not particularly surprising that this extends to all planets.

But *Kepler's first law* is something new. The statement of it is that the line from the Sun to the planet sweeps out equal areas in equal times.

Put another way, Kepler's first law tells us that, if we look at any triangle whose corners are the Sun and two points on the orbit near to each other, then the area of the triangle divided by the time it takes the planet to go from one point to the other will be the same, no matter where the points on the orbit are.

I have calculated this ratio for all pairs of adjacent points on the orbit shown in Figure 4.3. The resulting values are graphed in Figure 4.4. The values are constant, verifying Kepler's first law.

The first law contains important information. Given that we know the orbit of the planet, the law of areas allows us to calculate where the planet is after any time. It is the law that determines the speed of the planet in its orbit. Physicists and astronomers today have a different name for this law. It is called the law of **conservation of angular momentum**. The quantity physicists call the **angular momentum** of the body is just twice the mass of the body times Kepler's area sweeping rate (area swept out per unit time). The constancy of Kepler's area rate implies that the angular momentum is constant (which, in this case, is what physicists mean by "conserved").

The Sun has a little orbit of its own

In this section: the masses of the planets move the Sun.

As we remarked above, the equality of action and reaction means that if the Sun exerts forces on the planets, then the planets exert forces on the Sun, and the Sun must move in response to them. This motion turns out to be fairly small, but not insignificant. Because Jupiter is so massive, it exerts the largest force on the Sun.

Just as Jupiter moves on (roughly) a circle, so will the Sun. These circles should have a common center on the line joining the two bodies. The Sun's acceleration will be smaller than Jupiter's by the ratio of their masses (since the forces on the two are the same), and it must go around its circle with the same period as Jupiter's

orbit. We saw earlier that the acceleration of circular motion is proportional to the radius divided by the square of the period, so we can conclude that the radius of the Sun's circle is smaller than Jupiter's by the ratio of their masses. This ratio is 0.0009547, so the radius of the Sun's orbit is 743 100 km. For perspective, compare this to the Sun's own physical radius of 695 990 km.

Therefore the Sun executes an orbit about a point just outside itself with a period of 11.86 years. On top of this are smaller motions due to the other planets, particularly Saturn.

Geostationary satellites

Communications satellites relaying telephone and television signals from place to place on the Earth have to stay above their receiving and transmitting stations all the time in order to be effective. This means they have to be in an orbit which has a 24 hour period, so that the Earth will turn at just the right rate to keep their stations under them. Let us find out how this can be arranged.

We have seen in Chapter 3 that an orbit at an altitude of 300 km (a radius of 6700 km) has a period of about 90 min. By Kepler's third law, the square of the period is proportional to the cube of the radius of the orbit. Since we want a period that is roughly 16 times as long, we want the square of the period to be 256 times as large. The radius will depend on the cube root of this, which is 6.35: the radius needs to be 6.35 times as large as the radius of the 90-minute satellite. This is 6.6 times the radius of the Earth itself, or 42 500 km. This places such a satellite 36 100 km above the ground.

The gravitational attraction of spherical objects

There is one point in our deduction of the law of gravity that we have glossed over, but which caused Newton the most difficulty of all. The gravitational attraction exerted by a body falls off as the square of the distance to it, but if the body is not very small, what do we mean by the distance to it? When we calculated the Moon's acceleration, we assumed (with Newton) that the important distance was that to the Earth's *center*. The reason that Newton worried about this was that it wasn't just a matter of definition in his law, but rather a question of the self-consistency of his theory.

To understand what this means, consider the Earth not as a single body, but as a composite made up of tiny particles distributed throughout its whole volume. Each of these particles exerts a gravitational attraction on the Moon that is directed towards the particle itself, not towards the center of the Earth. Thus, the center of the Earth must be some *average* center of attraction. This is something Newton felt he would have to be able to deduce from his theory, rather than simply postulating it.

When he finally solved the problem, he found that the force of gravity does indeed vary as $1/R^2$ outside an exactly spherical body, where R is the distance to its center, but that the gravitational attraction of bodies with other shapes was more complicated. Now, since the Earth and all the other bodies in the Solar System are roughly spherical, his law of gravity could be used without significant error. It is an illustration of Newton's intellectual thoroughness and honesty that he was nevertheless unwilling to publish his theory of gravity until he had solved this subtle and difficult point.

The importance of this result goes far beyond the fact that it makes orbital calculations easier. It means that if I am at a given distance r from the Earth, then the size of the Earth does not affect the gravitational force I feel, provided that in changing the radius R of the Earth I do not change its mass, and provided that R

In this section: to arrange for an orbiting satellite to hover over the same location on the Earth, the satellite must be much further away than most.

In this section: a key to keeping orbit calculations simple is that spherical bodies in Newton's theory of gravity exert the same force on distant objects as they would if all their mass were concentrated at a point. The size of the Sun, for example, does not need to be taken into account when finding the orbits of planets. We prove this property using a computer program.

stays smaller than r . So if I were to imagine the Earth shrinking for some reason, it would not affect me if I stayed at the given distance r outside it. On the other hand, if I attach myself to the shrinking Earth's surface, then the gravitational acceleration at its surface will increase as $1/R^2$.

This difference between the behavior of gravity at different places is crucial to an understanding of *black holes*: when a star shrinks to form a black hole, gravity on its surface gets stronger and stronger, but at a fixed point outside the gravitational field does not change. We will return to black holes later on in this chapter.

We shall show that spherical bodies have the acceleration assumed above by the same method that Newton used, except that where he did his calculation using calculus, we shall do it on a computer.

Rather than deal with a whole sphere, it is sufficient to consider only a very thin spherical shell of matter, because the whole sphere can be built up out of such shells. Our shell will be subdivided into many tiny parts, each of them effectively a **point mass** at a different distance from the place where we want to compute the attraction. We shall show that by adding up all these separate attractions, we get a result which is proportional to the total mass of the shell and inversely proportional to the square of the distance to its center. The calculation is in Investigation 4.3.

The result of the computer calculation of Newton's result is shown in Table 4.4. In the first column is the number of zones into which I have divided the shell. Since each zone is treated as a point mass, we should expect that the calculation will get more accurate as the size of a zone shrinks, that is as the number of zones increases. The table bears this out, since the difference between the numerical result and Newton's gets smaller as the number of zones increases.

Other features of Table 4.4 are also worth noticing. Consider how the accuracy of the computer calculation with a fixed number of zones gets worse as the place where the force of gravity acts nearer the surface of the shell. This is an effect of the zoning: those zones nearest the point where we calculate the force make a big contribution to the force, since the force is proportional to the reciprocal of the distance squared. The fact that these nearby zones are treated as point particles when they really are not is more important if these zones are nearby. Nevertheless, when the number of zones is increased, the accuracy improves.

Another feature that Table 4.4 reveals is that the force of gravity *inside* the hollow sphere is zero! Just outside the shell the force is large, but after we cross inside the shell it drops to zero.

Newton was also able to show this. It means that if we dig a deep hole into the Earth (which we idealize as spherical for this discussion), the force of gravity that we feel depends only on the mass of the part of the Earth that is *inside* the radius that we have reached. The material outside our radius exerts no net gravitational pull on us. If we reach the center of the Earth (hypothetically!), the force of gravity will vanish entirely.

Playing with the orbit program

Having constructed the orbit calculator, we don't have to stop after just calculating a few planetary orbits. We shall use it below to calculate the Newtonian prediction of the deflection of light as it passes the Sun, which is responsible for the phenomenon of **gravitational lensing**. Another interesting question to ask is, what is the speed a planet needs in order to escape from the Sun, i.e. to get into an orbit that never comes back?

In this section: we experiment with different kinds of orbits.

Investigation 4.3. Spheres are just as attractive as point masses

A sphere can be decomposed into thin concentric spherical shells. If we can show the result for a thin shell then it will be true for a sphere, since all the shells have the same center. Suppose we have a shell made of a material with density ρ and of thickness ϵ . (The Greek letter ρ (rho) is the usual symbol physicists use for density. Physicists and mathematicians also use ϵ (epsilon) to represent something that is taken to be very small.) Choose any point on it as a North pole and put down lines of **latitude** θ and **longitude** ϕ . We shall use these lines to form a grid on the sphere. We shall take any small section of the sphere thus marked out and idealize it as a point mass. By adding up the gravitational forces of these point masses we will get Newton's result.

Suppose the angles θ and ϕ are measured in degrees, some of which are shown in Figure 4.5 on the following page. Suppose further that the grid of lines of latitude and longitude are spaced apart by the small angles $\delta\theta$ and $\delta\phi$, respectively. If the radius of the Earth is R , then the circumference of the circle of constant latitude shown in Figure 4.5 is $2\pi R \cos \theta$. Any small segment of it of angular length $\delta\theta$ will have a length in proportion: $2\pi R \cos \theta (\delta\theta/360)$. A line of constant longitude ϕ is a great circle of circumference $2\pi R$, so a small segment of angular length $\delta\phi$ has length $2\pi R (\delta\phi/360)$. So any small region of the sphere enclosed by pairs of adjacent grid lines, as in Figure 4.6 on the next page, has an area equal to the product of these,

$$\text{area} = \left(\frac{2\pi R}{360} \right)^2 \cos \theta \delta\theta \delta\phi. \quad (4.11)$$

This is really only an approximate answer for the area, because we have used the formula for the area of a rectangle in a plane, and the region is really part of a sphere. However, provided $\delta\theta$ and $\delta\phi$ are small enough, the error won't be large.

Now, since the shell has thickness ϵ , the volume of the tiny region we will approximate as a point mass is ϵ times the area, and so its

mass is ρ times this,

$$\text{mass} = \epsilon \rho \times \text{area}.$$

The computer program to do the calculation is called Sphere-Gravity, on the in website. It has no special tricks. After choosing both a location at which we want to calculate the acceleration of gravity and a radius for the shell, we just calculate the acceleration due to each piece of the shell and add them all up. For convenience we take the point at which we want to compute the acceleration to be at the origin of our x - y - z coordinate system, and we take the sphere to be centered at the point a distance d from the origin on the x -axis. (See Figure 4.6 on the following page.)

If the North pole is also on the x -axis, a distance $d + R$ from the origin, then a point on the shell at latitude θ is a distance $R \cos \theta$ from the x -axis and a distance $d + R \sin \theta$ along the x -axis, so it is at a distance $r = (d^2 + R^2 + 2Rd \sin \theta)^{1/2}$ from the origin. (Note that we call θ here the latitude angle that was called β in Figure 4.5 on the next page.) The acceleration this small section of the shell produces in the x -direction is the mass of the small piece of the shell divided by r^2 times the cosine of the angle α in the diagram. This is just $(d + R \sin \theta)/r$. The computer program simply multiplies the mass of the piece by the cosine factor and divides by the square of the distance. It then adds up all these contributions from each of the little patches on the sphere. (Notice that I have left out the constant G . It is not important for this calculation: we want to show that the sum of the accelerations of the small pieces equals Newton's acceleration, and this will still be true if we divide both by G .)

Each piece of the shell also produces accelerations perpendicular to the x -axis, but these must sum to zero because of the symmetry of the problem. Since each direction perpendicular to x is equivalent to every other one, the net attraction could not point along any one perpendicular direction, so it points along none: its perpendicular component must be zero. So we need not bother to add these up.

Exercise 4.3.1: Area of Colorado

The American state of Colorado is a spherical rectangle of the kind we have just described. Its northern and southern boundaries have latitude 41° and 37° , respectively. Its eastern and western boundaries have longitude 102° and approximately 109.1° , respectively. Given that the radius of the Earth is 6.3782×10^6 m, what is the area of Colorado?

We shall show in Chapter 6 that the escape speed is $(2GM/R)^{1/2}$, where R is the radial distance from the Sun at which the planet starts out. This is just $\sqrt{2}$ times the circular orbital speed, in other words 41% larger. However, this does not depend on the initial direction the planet takes, as long as it doesn't actually crash into the Sun. You might like to try to use the orbit program to test this equation and its lack of directional dependence.

Here again the test cannot be perfect, not only due to the numerical errors but also because no one can allow the program to calculate forever in order to verify that

Number of patches	Ratio of radius of shell to distance to its center					
	0.01	0.1	0.5	0.9	1.5 ^a	5.0 ^a
400	0.10	0.11	0.23	6.26	0.20	0.002
2500	0.016	0.01	0.037	0.93	0.032	0.0003
40 000				0.052		

^aThe figures given in the last two columns are the computed acceleration as a percentage of the acceleration that there would be at that point if the shell were small enough to be inside that radius.

Table 4.4. Gravitational attraction of a spherical shell. Figures given are percentage deviation of the numerical result from the exact Newtonian result. In the first four columns, increasing the number of divisions of the shell clearly brings the numerical answer closer to the Newtonian one. In the last two columns the acceleration is calculated at a point inside the shell, where the Newtonian answer is zero.

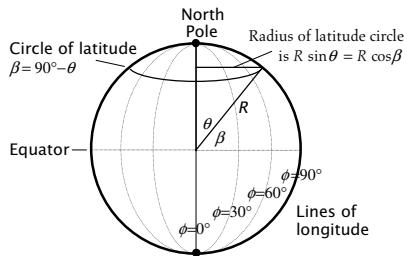


Figure 4.5. A grid of lines of longitude and latitude form a coordinate system for a sphere, as used for the calculation in Investigation 4.3 on the previous page. A circle of constant latitude has a radius that depends on where it is. Circles of constant longitude all have the same radius, that of the sphere itself.

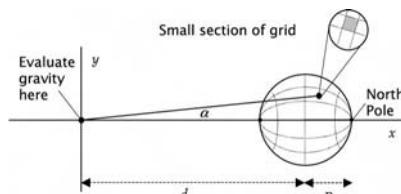


Figure 4.6. Geometry for the calculation of the force of a spherical shell in Investigation 4.3 on the previous page. The force is calculated at a location a distance d from the center of the sphere, whose radius is R . Each small section of the sphere, like the one shown, makes a contribution.

the planet doesn't really turn around at some large distance. Nevertheless, it should be possible to get a reasonably good demonstration with only a few calculations. Naturally, one can also calculate the escape speed from the Earth if one uses the mass of the Earth in the formula. For a space probe launched from the Earth's surface, this speed is 11 km s^{-1} .

The orbit program might tempt those with lots of computer time to try to evolve the whole Solar System for a long period of time, but this is not recommended! Inaccuracies build up after a few dozen or a few hundred orbits and render results over longer spans meaningless. Faster computers can run more accurate programs, but still the results cannot be trusted for more than a few tens of millions of years. This is one reason that, despite our knowledge of the "exact" laws of planetary motion, we have a very incomplete knowledge of what the early Solar System was like.

Black holes before 1800

In this section: the equivalence principle and the realization that light has a finite speed led 18th century physicists to the possibility of "dark stars", stars that exert gravity but cannot emit light. These are the Newtonian versions of Einstein's black holes. Not until long after Einstein did scientists realize that Nature creates these objects abundantly.

▷ Although Laplace is well-known to most modern physicists and mathematicians, Michell has sunk into obscurity. This is somewhat unfair, since in his day he was regarded as one of the premier scientists in Britain. It was he who suggested to his friend Cavendish the experiment to measure G that we referred to above.

▷ Today we know c has the value $2.998 \times 10^8 \text{ m s}^{-1}$, quoting only the first four figures.

We now have enough knowledge about gravity to take a look at one of the most remarkable speculations of eighteenth century physics: what we now call the black hole. In the late 1700s the British physicist John Michell (1724–1793) and the French mathematician and physicist Pierre Laplace (1749–1827), both of whom were well-acquainted with Newton's laws of motion and gravity and with the equivalence principle, independently put together two simple facts:

1. No object can escape from a body if its speed is less than $(2GM/R)^{1/2}$.
2. Light travels at a finite speed c . This had been proven by the Danish astronomer Olaf Roemer (1644–1710) in Newton's time, but the value of c was not well-known.

They then reasoned that light cannot escape from a body whose escape speed exceeds c . This inequality can be solved for the radius R of the body to give

$$R \leq R_g , \quad \text{where} \quad R_g = \frac{2GM}{c^2}. \quad (4.12)$$

So if it were possible to shrink a body of a fixed mass M down to a size smaller than R_g then it would appear black to the outside world.

Notice that the limiting radius R_g depends only on M and on the constants of nature c and G . Today we call this the *gravitational radius*

of a body of mass M . The remarkable thing is that it gives exactly the radius of what general relativity calls a black hole, which is something from which no light can escape.

We shall look in more detail at the modern notion of a black hole in Chapter 21, but here it is worth saying that the Michell–Laplace black hole is not identical to its modern counterpart. In particular, Michell and Laplace envisioned light *starting off* with speed c as it leaves the body and gradually slowing down as it gets further away, eventually actually turning around and falling back in if the inequality was satisfied. So someone close to the body might still see a few “tired” beams of light before they turned around. Today, however, we believe that light always travels at speed c , and that if it leaves a body it cannot then turn around and fall back in. The modern idea of a black hole is that if a body has a radius smaller than its gravitational radius, then light never leaves it at all. No observer anywhere outside it would see any light from it at all.

Nevertheless, Michell’s and Laplace’s fundamental instincts here were sound. They had the courage to extend the equivalence principle to light, to say that light was affected by gravity just the same as anything else. The equivalence principle is fundamental to Einstein’s general relativity, and underlies the modern black hole as well. Michell and Laplace could not have anticipated the development of relativity more than a century later, but within their own perspective they showed remarkable vision.

Light is deflected by the Sun's gravity

Another consequence of the equivalence principle that nineteenth century physicists were perceptive enough to work out is that light will change direction as it passes the Sun. The reason is the same as the one underlying black holes: the effect of gravity on a particle depends only on the particle’s speed, so if we set that speed to c then we will find out what happens to light itself.

We can do this by simply adapting our computer program `Orbit` that calculates Solar System orbits. Instead of using initial conditions appropriate to Mercury or another planet, we use initial conditions for a light ray coming from a distant star. The light will initially be traveling on a straight line at speed c . Its speed is much larger than the escape speed of the Sun, so that after passing the Sun, it is again moving on a straight line; but its direction will be different. Running the computer program `Orbit` with three different sets of initial conditions of this kind leads to the trajectories shown in Figure 4.7 on the following page.

To understand the diagram, we need to discuss the size of the expected deflection. Suppose, if the light were not affected by gravity, that the line would pass a minimum distance d from the center of the Sun. This is called the *impact parameter* of the light ray. In Investigation 4.4 on the following page we show that the deflection angle, measured in **radians**, is roughly $2GM/c^2d$, where M is the mass of the Sun. A light ray just grazing the surface of the Sun would have an impact parameter approximately equal to the radius of the Sun, $d = 7 \times 10^8$ m, from which we would deduce a deflection of less than one second of arc (less than 10^{-6} radians). This would not be noticeable if plotted on a graph like Figure 4.7, so to do the figure I artificially shrunk the Sun to a point and allowed the light to have a very small impact parameter. I chose three values of d : 10 km, 15 km, and 20 km. The deflections measured from the graph are, respectively, 16.9° , 11.3° , and 8.4° . It is easy to check that they agree with the prediction of the above formula, which we calculated only roughly in Investigation 4.4 on the next page.

This formula for the deflection was first derived by Cavendish in 1784 and inde-

In this section: another “modern” phenomenon that was anticipated long ago by physicists working in Newtonian gravity is the fact that light rays, when passing close to the Sun, alter their direction. Einstein’s theory was not new in making this prediction, but it predicts twice as large an effect as Newtonian theory. This was first verified in 1919.

Investigation 4.4. The Newtonian deflection of light

We shall derive the deflection by using the principle of equivalence, in much the same style as we derived the gravitational redshift in Investigation 2.2 on page 16. Consider light passing a star. Since light must travel on a straight line with respect to a local freely-falling observer, and since these observers all fall towards the center of the star, the light must continually bend its direction of travel in order to go on a straight line with respect to each observer it happens to pass. We can estimate the size of the effect, at least roughly, by the following argument within Newtonian gravity.

Let us consider just one freely-falling observer, who is at rest with respect to a star of mass M at the point where the light beam makes its closest approach to the star as it passes it by. Let this closest distance be d , the impact parameter. (This is not quite how we defined the impact parameter, but it is close enough for our approximate argument.) The observer's acceleration towards the star is $\mathbf{g} = GM/d^2$.

Traveling at speed c , the beam of light will experience most of its deflection in a time of order d/c , the time it takes for the light to move significantly further away from the star. During this time, the observer has acquired a speed $v = gd/c = GM/cd$ perpendicular to the motion of the light. By the equivalence principle, the light must also have acquired roughly this same speed transverse to its original direction. Since its speed in the original direction is very little changed, we can calculate the angle of deflection by simple geometry: the tangent of the deflection angle is v/c . For small angles, the tangent of an angle is equal to the angle as measured in radians. This leads to the estimate that the angle of deflection will be

$$\alpha = v/c = GM/c^2 d \text{ radians.}$$

The total deflection should be double this, since the light will experience the same deflection coming in to the point of nearest approach as going out:

$$\text{Newtonian prediction of the deflection angle in radians} = \frac{2GM}{c^2 d}. \quad (4.13)$$

This turns out to be *exactly* the answer that a very careful calculation would give. But we do not need to do that calculation to check Equation 4.13. We only need to use the computer program *Orbit* with the right initial data and measure the results.

How would one measure this? If we look at the position of a star just once, as its light is passing near the Sun, we won't know how much deflection it is suffering, since we don't know its true position. The way to do it is to measure the position of a star when its light is passing nowhere near the Sun. Then, perhaps six months later, when the Sun is near the position of the star, measure the position again. It is clear from Figure 4.7 that the apparent position of the star moves *outwards*, away from the Sun. The only difficulty is in seeing the star when the Sun is near. But during an eclipse of the Sun, all the light from the Sun is blocked by the Moon, and so it is possible to see stars very close to the Sun's position. This is how the effect was eventually measured. The observed result is twice the number given by Equation 4.13, consistent with general relativity.

Exercise 4.4.1: Light deflection by other bodies

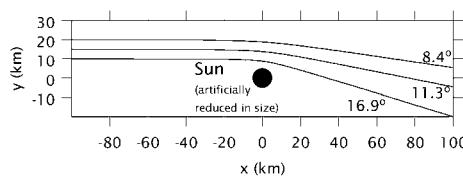
Any gravitating body will deflect light. Estimate, using Equation 4.13 above, the amount of deflection experienced by a light ray just grazing the surface of the following bodies: (a) Jupiter, whose radius is 7.1×10^4 km; (b) the Earth; (c) a black hole of any mass; and (d) you.

pendently by the German astronomer Johann G von Soldner (1776–1833) in 1801. They regarded it as a mere curiosity, since measuring the apparent positions of stars to an accuracy of an arcsecond or so was impossible in their time. Einstein himself independently re-derived the formula using the equivalence principle in 1909, and he pointed out that the expected deflection might well be observable with the telescopes of his day.

Figure 4.7. The deflection of light by a point-like mass as calculated in Newtonian gravity. The mass has the mass of the Sun, but to show the effect, it has been made almost as compact as a black hole.

This allows trajectories to experience strong gravity and exhibit large deflections.

► Eddington was one of the first true *astrophysicists*, a scientist who used the theories of physics to understand the nature of astronomical objects. He was an early champion of Einstein's general relativity. Dyson was the British Astronomer Royal from 1910 to 1933.



However, before a suitable opportunity arose for observing the effect, Einstein moved on to devise the theory of general relativity (1915). In this theory, there is an extra effect that causes light to deflect twice as much, so that the new prediction would be $4GM/c^2 d$. (We shall calculate this effect in Chapter 18.) A deflection of this size was indeed measured in an eclipse expedition in 1919 led by the British astronomers Sir Arthur Eddington (1882–1944) and Frank W Dyson (1868–1939). The accuracy was enough to distinguish between the old Newtonian deflection and the new general relativistic one. The verification of this prediction of general relativity did more than anything else to make Einstein a celebrity, a household name.

Tides and tidal forces: the real signature of gravity

The tides wash the margins of all the great oceans, regulate the lives of sea urchins and fishermen, power the great **bore waves** on rivers like the St. John, the Amazon, and the Severn. For most of us the tides are romantic, primeval, poetic. Standing on an ocean beach, we might be impressed by this tangible manifestation of the gravity of the distant rock we call the Moon, but few of us would be led to reflect on how fundamental the tides are to an understanding of gravity itself. But *fundamental* is the right word. In the modern view, the *real* signature of gravity, the part of gravity that can't be removed by going into free fall, is the *tidal force*, whose most spectacular effect on Earth is to raise the ocean tides. In this chapter we will examine this aspect of gravity, starting with the simplest effects first and working our way up to ocean tides and then to tides elsewhere in the Solar System and beyond. We will return again and again in later chapters to the fundamental role of tides. Indeed, many astronomical systems transmit tidal forces as signals right across the Universe, signals that we call **gravitational waves**.

Tidal forces in free fall

When we formulated the modern version of the equivalence principle in Chapter 2, we talked about experiments performed in free fall. The simplest such experiment is just to carry a stone with us in free fall and then release it at rest. The principle of equivalence says that nothing happens: it just stays alongside us as we fall. A little thought will convince us that this is only true if the body is very near to us, and if we limit the duration of the experiment.

Consider two stones falling freely towards the Earth, one just above the other. The lower stone, being closer to the Earth, experiences a slightly larger acceleration of gravity than the higher stone, since gravity gets weaker at larger distances. This means that even if the two stones start out falling with the same speed, the lower one gradually acquires a slightly larger speed than the higher one, and the distance between them increases, as in Figure 5.1 on the following page.

This is due entirely to the fact that the Earth's gravitational field is *non-uniform*: it pulls with different accelerations in different places. If the acceleration of gravity were strictly the same everywhere, then the two stones would stay at the same separation forever. The non-uniformity (we sometimes say **inhomogeneity**) of the Earth's gravitational field has the effect of pushing the stones apart if they are placed one above the other.

Next consider two stones falling side by side. If they start from rest, they will both fall on radial lines directly toward the center of the Earth. This means they will not quite keep their initial sideways separation: they will approach each other as they approach the Earth. Again this is an effect of the non-uniformity of the Earth's gravitational field, which pulls in different directions at different locations.

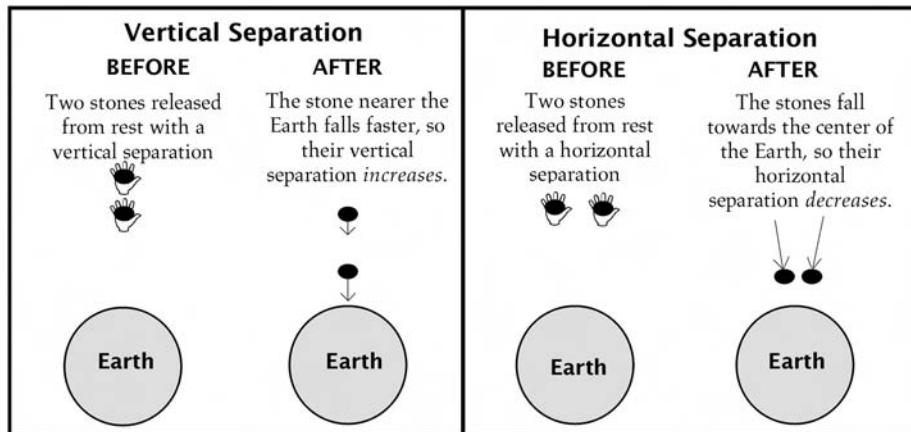
To an experimenter falling freely with the stones, the overall acceleration of gravity disappears (the equivalence principle), but these residual

In this chapter: we study tidal gravitational forces. These are the forces that are not removed in free fall, because they come from non-uniformities in the gravitational acceleration. Their effects are visible all over the Universe, from the ocean tides on the Earth to the disruption of whole galaxies when they get too near to one another. The precise calculation of the tidal effects on Mercury's orbit left a tiny part of Mercury's motion unexplained by Newtonian gravity, its first failure. Einstein's general relativity explained the discrepancy.

In this section: tides arise from non-uniformities in gravitational accelerations. They are the part of the gravitational field that cannot be eliminated by going into free fall.

►The figure on this page shows a volcanic eruption on Io, imaged by the Voyager 2 spacecraft in July 1979. Such eruptions are frequent, and are the result of the heating of the moon as it is deformed by changing tidal forces (see later in this chapter). Image courtesy of NASA/JPL-Caltech.

Figure 5.1. Tidal effects are most easily seen in the motion of freely-falling objects.



tidal effects of gravity remain: stones placed one above the other are pushed apart, while stones placed side by side are pulled together.

Notice that the tidal effect is proportional to the distance separating the stones, at least for relatively small separations. Horizontally separated stones accelerate toward each other in proportion to their separation: the larger their distance, the greater their **tidal acceleration**. The same is true for the vertically separated stones. This aspect of the tidal force suggests that it could be a significant force on scales of the diameter of the Earth, even though we don't notice it in local experiments. We shall also find that the increase of tidal forces with separation will be important to our discussion of the detection of gravitational waves, in Chapter 22.

The discussion of tidal forces we have just given would certainly have been acceptable, even obvious, to a nineteenth century physicist. But he might not have been prepared to place the tidal forces on a pedestal and call them the "real" gravitational force, the way we do today. It was principally Einstein who stressed the fundamental importance of the equivalence principle, who made uniform gravitational fields seem trivial, and who gave the tidal forces a special mathematical place in his theory of gravity, general relativity. We shall adopt this modern perspective here, and pay due respect to tidal forces by devoting this chapter to them.

But what about the equivalence principle? Are tidal forces its downfall? No, but only if we keep it *local*: given an experimenter in free fall who can measure things (such as distances, speeds, etc.) only to a certain accuracy, then there will be a certain region of space around him and a maximum duration of time for experiments in which he will not be able to detect the effects of the tidal forces. In this region the equivalence principle will be valid. Since no measurement is perfectly accurate, there is a real sense in which the equivalence principle applies in a small region but not everywhere.

Physicists use the words **local** for things that apply in small regions and **global** for things that apply everywhere. We therefore say that the equivalence principle is valid locally but not globally.

Ocean tides

In this section: how tides work; the way the Moon raises tides in the Earth's oceans.

Let us now see how tidal forces actually raise the tides. The Moon exerts a gravitational force on the Earth, equal and opposite to that which the Earth exerts on it. In response to this force, the Earth executes a small circular motion about its average orbital motion as it circles the Sun, just as the Sun orbits a point near it because of

Jupiter's force on it. (Recall the discussion of this motion in Chapter 4.) But this small circular motion is not the whole story, because the force which any piece of the Earth feels from the Moon's gravity depends on how close it is to the Moon.

The Earth is a large body, so the acceleration of the Moon's gravity on the side of it nearest the Moon can be substantially larger than on its far side. If the Earth were made of tissue paper, this difference in acceleration would probably tear it apart. But the Earth is tougher than that: its internal forces (both the mechanical forces that make rocks rigid and its own gravitational force on things on its surface, like oceans and people) are more than strong enough to resist this difference in acceleration.

Because of these internal forces, the parts of the Earth nearest the Moon cannot fall freely in the Moon's gravitational field, so they do not accelerate with the full acceleration of the Moon's gravity there. Instead, they stay attached to the Earth and accelerate at the same rate as the rest of the Earth. Similarly, parts of the Earth furthest from the Moon would, if they were free of the Earth's forces, fall toward the Moon with a smaller acceleration than the rest of the Earth, but they are not free to do so: they stay attached to the Earth, too. The net effect is that the Earth accelerates toward the Moon with the *average* of the acceleration of the Moon's gravity across it. Only the *center* of the Earth is truly freely-falling. This is illustrated in Figure 5.2.

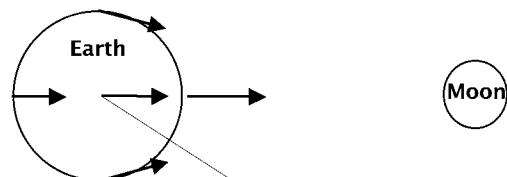
Now consider an experimenter sitting at the center of the Earth. From his freely-falling point of view, the vertical stretching action of the tidal force of the Moon's gravity will try to pull the side of the Earth nearest the Moon away from the center. Even though the Earth's internal forces hold it together against this pull, the Earth is not perfectly rigid, and it will bulge slightly toward the Moon. The tidal force has an even more drastic effect on the oceans, which are not rigidly connected to the surface. An ocean whose center is on the side facing the Moon will be raised in elevation by this force, causing it to pull away from its shores, giving a low tide. An ocean whose edge is on the side nearest the Moon will find its water drawn towards this edge, giving a high tide.

But the part of the tidal force that might be more unexpected is that this *same* thing happens on the side of the Earth furthest from the Moon as well. Again as seen by the freely-falling experimenter at the center of the Earth, the vertically stretching tidal force of the Moon pushes the far side of the Earth *away*. What is really happening here of course is that the more weakly accelerated far side of the Earth is being left behind as the Earth accelerates toward the Moon: it is pulled toward the Moon, but not as strongly as the Earth as a whole, and so, *relative to the center of the Earth*, it is pushed away. An ocean centered on the far side will bulge out, too, and its shores will experience a low tide as well.

In one day any given ocean will experience *two* low tides, once because of its nearest approach to the Moon and the other because of its farthest recession away from it. Tidal effects have this characteristic behavior under rotations: places where the tidal effect is similar are separated by a rotation of only 180°.

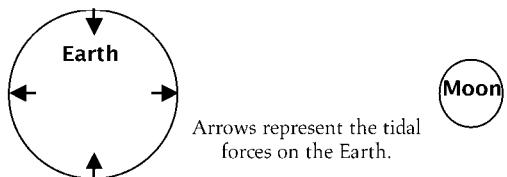
We will encounter this symmetry again when we discuss gravitational waves in

Acceleration of the Moon's gravity on Earth.
Length of arrow indicates size of acceleration.



The acceleration at the **center** is the mean acceleration with which the solid Earth will fall. The acceleration of gravity due to the Moon is larger near the Moon and smaller further away.

Residual acceleration of the Moon's gravity,
after subtracting the mean acceleration of the Earth.



Arrows represent the tidal forces on the Earth.

Figure 5.2. Tides raised by the Moon on the Earth arise from the residual acceleration of gravity left when the solid Earth falls at the average of the acceleration of the Moon's gravity across the Earth.

Chapter 22.

Actually, the first high tides of consecutive days do not occur at exactly the same time of day, because the Moon has moved along in its orbit during the intervening day. Since the Moon orbits the Earth in the same direction as the Earth turns, any place on Earth has to wait somewhat more than 24 hours before it is again closest to the Moon. The Moon takes 27.3 days to orbit the Earth, so in one day it moves a fraction $1/27.3$ of an orbit. Since the Earth takes 24 hours to turn full circle, it takes a fraction $24/27.3$ of an hour to turn through the same angle as the Moon goes through in a day. This amount of time, about 53 min, is the amount by which a high tide is delayed past the time it arrived on the previous day. Similarly, the time between successive high (or low) tides is half of the full day-to-day period, about 12 hours 26 min.

Tides from the Sun

In this section: the Sun raises tides almost as high as the Moon does.

The Moon is not the only body strong enough to raise tides on the Earth. The Sun, though much further away, is also much more massive, and its tidal forces on the Earth are very similar in size to those of the Moon. The other planets have a negligible effect.

The fact that the Sun and Moon happen to exert similar tidal forces on the Earth is deeply related to another “accidental” fact that might at first seem to be completely unconnected, namely that **eclipses** occur, i.e. that the Moon and the Sun are of similar angular size on the sky.

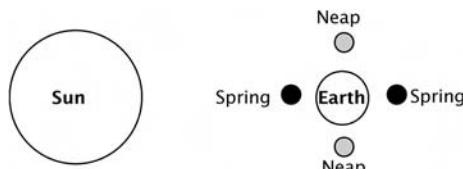
The reason for the relation between these two facts is explored in Investigation 5.1.

Spring and neap tides

In this section: the action of the Sun and Moon together is responsible for the seasonal variations in tides, from spring to neap and back again.

Investigation 5.1 shows that the Sun has a tidal effect on the Earth that is about 42% of that of the Moon. Thus, when these two effects reinforce each other, the tidal forces are 1.42 times the Moon’s alone, while when they work against each other they are only 0.56 times the Moon’s. Thus, the ratio of the maximum to the minimum tidal force is about 2.5.

Figure 5.3. Orbital alignments that give spring tides and neap tides. When the Moon is at the location of the darkly shaded circles, the tidal forces on the Earth are at their maximum. When the Moon is at the location of the lightly shaded circles, the tidal forces are at their minimum.



fects occur when the three bodies form a right triangle. This is a “neap” tide. (The word “neap” comes from Old English, where it meant helpless or weak.)

Therefore, there are two spring tides per month and two neap tides, and the spring tides are associated with the full and new Moons.

For the same reason that the interval between successive high tides in a day is slightly more than 12 hours, so too the interval between successive spring tides is slightly more than half of the Moon’s orbital period, which is 13.7 days. Since the line joining the Sun and the Earth has rotated in this time because of the orbital motion of the Earth, the actual time it takes for the Moon to rejoin this line is a little more than a day longer, 14.8 days.

When do the two forces add? Since the tidal effects are the same on opposite sides of the Earth, the tidal forces of the Sun and the Moon reinforce each other if the three bodies all lie on a straight line, either with the Moon between the Earth and the Sun or on the other side of the Earth, as in Figure 5.3. This tide is called a “spring” tide. Similarly, the minimum tidal ef-

Investigation 5.1. Tides and eclipses

We want to calculate the relative strength of the Sun's and the Moon's tidal forces on the Earth, and to do this we have to discover how the tidal force depends on the distance r of the Earth from the body producing the tides. The overall gravitational force falls off as $1/r^2$, but the tidal force, which is the difference between the forces at two nearby points, falls off faster, as $1/r^3$.

To see this, we introduce an important algebraic expression, called the binomial theorem. This gives the value of $a + b$ raised to any power n , where a and b are any two numbers:

$$(a + b)^n = a^n + nba^{n-1} + \frac{1}{2}n(n-1)b^2a^{n-2} + \dots, \quad (5.1)$$

where I have only given the first three terms. If n is an integer, only the first $n+1$ terms are non-zero. If $n = 1$ the third term is zero, and we have the simple identity $(a + b)^1 = a + b$. If $n = 2$ then we have the quadratic formula $(a + b)^2 = a^2 + 2ab + b^2$, and the terms represented by the \dots in Equation 5.1 all vanish. If $n \geq 3$, Equation 5.1 only gives the first three terms. However, there is an important special case where the extra terms left out don't affect the answer very much. Consider the form of Equation 5.1 when $a = 1$:

$$(1 + b)^n = 1 + nb + \frac{1}{2}n(n-1)b^2 + \dots.$$

If in addition b is very small compared to 1, then b^2 is even smaller, and higher powers of b get smaller and smaller, so that even if they were considered in this equation they would not make much contribution. We have arrived at a very useful conclusion: letting ϵ denote the very small number, we have that if ϵ is sufficiently small, then

$$(1 + \epsilon)^n \approx 1 + n\epsilon. \quad (5.2)$$

(The symbol “ \approx ” stands for “is approximately equal to.”)

Consider two points a distance r and $r + h$ from the gravitating body, on the same radial line. The acceleration at distance r is k/r^2 , where k is a constant. The acceleration at the distance $r + h$ is

$$\frac{k}{(r+h)^2} = \frac{k}{r^2(1+h/r)^2},$$

where I have factored r out of the term in the denominator. By the binomial approximation, this is approximately given by

$$\frac{k}{r^2}(1 + h/r)^{-2}. \quad (5.3)$$

We can now use Equation 5.2 to evaluate the factor containing h/r in this expression. Since h is small compared to r (for the Moon and

Earth we have seen that h/r is about 1/60), terms containing $(h/r)^2$ (or higher powers of h/r) are small compared to the term involving h/r itself, and we will neglect them. The result is that

$$(1 + h/r)^{-2} = 1 - 2h/r + \dots,$$

so expression (5.3) becomes

$$k/r^2 - 2kh/r^3 + \dots. \quad (5.4)$$

This is the acceleration at $r + h$. The tidal acceleration is the difference between this and the acceleration at r itself, which just means subtracting off the first term of Equation 5.4. The result is

$$\text{tidal acceleration} = -2kh/r^3. \quad (5.5)$$

This establishes the $1/r^3$ fall-off of the tidal effects.

Let us now get rid of the constant k in the above expressions and replace it by what we know it to be, $-GM$, where M is the mass of the body producing the gravity. This tells us a crucial fact, that the tidal force is proportional to M/r^3 . Now, the average density ρ of a body is its mass divided by its volume. Since the volume of a sphere is proportional to the cube of its radius R , its density is proportional to M/R^3 . Turning this around, we find that its mass is proportional to ρR^3 .

This in turn means that the tidal forces it produces are proportional to $\rho(R/r)^3$. Now, the ratio $2R/r$ is the *angular diameter* of the sphere on the sky, measured in radians. Put another way, the number of degrees of arc that a sphere spans on the sky is 360° times the fraction of a full circle that its diameter occupies, which is proportional to R/r . Therefore the tidal acceleration is proportional just to $\rho \times (\text{angular diameter of the body})^3$, with a constant of proportionality that depends only on pure numbers (like π), Newton's constant G , and of course the difference in the positions of the two points whose tidal effects are being examined.

Since the Moon and the Sun have almost the same angular size as seen from the Earth (which is why eclipses are so spectacular), the tidal accelerations they produce across the diameter of the Earth are proportional to their densities. The Moon, made of rocks, has an average density of 3300 kg m^{-3} , which is about 2.4 times as dense as the Sun. Therefore the Moon's tidal effects on the Earth are 2.4 times as large as the Sun's. This makes the Sun less important than the Moon, but not of negligible influence. Other planets have similar densities to the Sun and Moon, but very much smaller angular diameters, so they do not exert significant tidal effects on the Earth.

Exercise 5.1.1: Testing the binomial approximation

Use a pocket calculator to verify that Equation 5.2 gives a good approximation for small ϵ . For the following values of ϵ and n , evaluate the approximate value $1 + n\epsilon$, the exact value $(1 + \epsilon)^n$, the error (their difference) and the *relative error* of the approximation, which is defined as the error divided by the exact value: (a) $n = 2$, $\epsilon = 0.01, 0.1, 1.0$; (b) $n = 3.5$, $\epsilon = 0.01, 0.1$; and (c) $n = -2$, $\epsilon = 0.01, 0.1$. (Recall that negative powers indicate the reciprocal, so that $(1 + \epsilon)^{-2} = 1/(1 + \epsilon)^2$.)

In the above discussion of how the two tidal forces add up, we have made the unspoken assumption that the orbit of the Moon around the Earth is in the same plane as the orbit of the Earth around the Sun, so that the three planets can actually form a straight line when they are in their best alignment. This is very nearly the case, but not quite. The Moon's orbit is inclined at an angle of 6° to the Earth's orbital plane, tilted in a direction that rotates with time (a period of roughly 20 years) because of the Sun's gravity. This means that twice a year the best alignment of the three bodies is as much as 6° away from a straight line, while three months later, when the line they form is nearly parallel to the intersection of the two planes, their alignment can be nearly perfect. This gives a small seasonal variation in the strength of the spring and neap tides.

In this section: the action of tidal forces results in a large variety of phenomena. The tides have locked the spin of the Moon to its orbital period, and they are gradually driving the Moon further and further from the Earth.

What the tidal forces do to the oceans, the Earth, and the Moon

Once Newton understood the universal nature of gravity and its variation with distance, he realized that he could explain the tides in the manner which we have just described. This was a significant piece of experimental support for his theory of gravity. But if we try to go further than this and actually predict the time of arrival of a tide at a particular place, the tidal range (difference in the height at high and low tides), and the variations of these with the day of the month, we find that the problem is hopelessly complex.

The reason is not hard to understand. Although the tidal *forces* are easy enough to describe, the response of the oceans to them depends on a large number of variables: the depth of the ocean in various places, the shape of the coastline on which the tides are measured, the density of the ocean (which depends on how much salt it contains), and even such day-to-day irregular conditions as the local atmospheric pressure and the strength of the winds. Even in the present age of modern computers, the prediction of tides along complicated coastlines is largely a matter of judgment based on their behavior in the past.

In certain places, such as the Bay of Fundy in Canada or the Bristol Channel in Britain, where the tidal surge of a large body of water is funneled into a narrow end at just the right distance from the opening, the tidal range can exceed 20 m and the pressure of this water can force a spectacular cresting wave that travels upstream in rivers that empty into this end. This wave, called a tidal bore (Figure 5.4), is very sensitive to changes in the tidal response of the body of water, so that even fluctuations in atmospheric pressure due to storms at sea can have a marked effect on the wave. The largest in the world is at Ch'ient'ang'kian (Hang-chou-fe) in China, where the wave can exceed 7 m in height.



Figure 5.4. The bore traveling upstream (to the left) along the River Severn in England in September 1976. The wave is breaking gently. Note the difference in water level before and after the wave. Enthusiasts often surf this wave. Photo copyright by the author.

The ocean tides are the most obvious manifestation of tidal forces, but what is the effect of the tides on the rocky body of the Earth and the Moon? Tides raised by the Earth have dramatically affected the Moon, causing it to show the same face to the Earth at all times. It is interesting to see how this happened.

At one time the Moon was spinning much faster on its axis than it is now.

The tides raised by the Earth would

make the rocks of the Moon bulge toward and away from the Earth, but the rotation of the Moon tended to carry this bulge around with it. Because friction slows down the response of any system to the forces on it, friction in the Moon caused the bulge in any group of rocks to lag behind the tidal force driving it. This meant that the bulge actually reached its maximum in the rocks shortly after they had rotated *past* the point of closest distance (and, on the far side, the point of furthest distance) to the Earth. The Moon therefore presented a bulge to the Earth that was not quite pointing towards the Earth. This bulge is illustrated in Figure 5.5, where its size and lag angle have been exaggerated for clarity.

The tidal force of the Earth naturally tried to align this bulge with the Earth-Moon line, which in this case meant pulling on the bulge *against* the direction of rotation (see Figure 5.5). The result was that the tidal force of the Earth slowed down the rotation of the Moon. As the rotation got slower and any group of rocks took more time to pass through the region nearest the Earth, the effect of the tidal

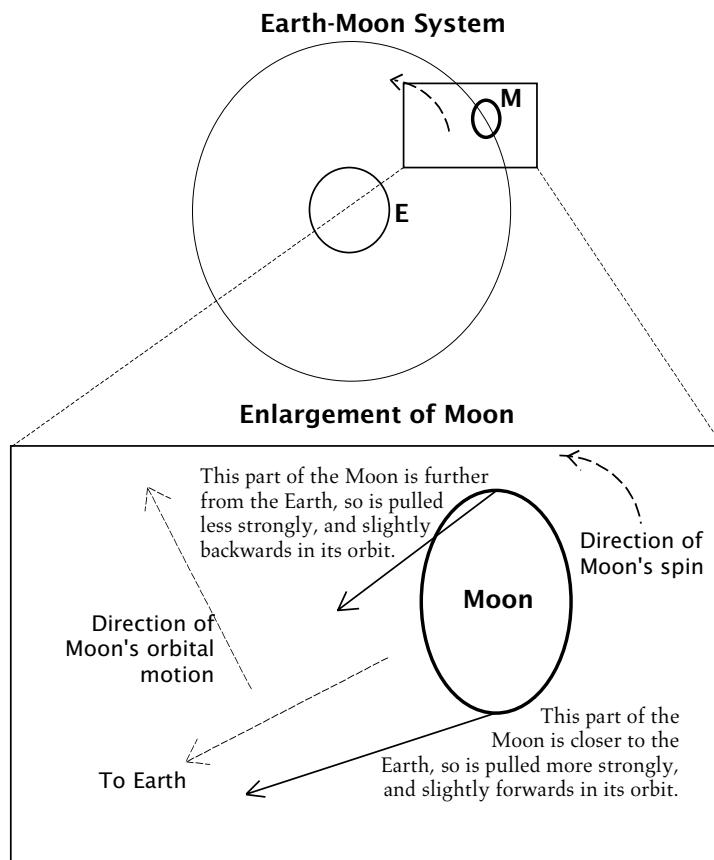


Figure 5.5. How tidal forces led the Moon to show the same face to the Earth at all times. The Moon's bulge lagged behind the tidal forces when it was spinning rapidly, leading to an elliptical shape that did not point towards the Earth. All rotations in this diagram are counterclockwise: the spin of the Moon and its orbital motion. The diagram illustrates the Earth's gravitational forces on the Moon at two places, on the near-side bulge and on the far-side bulge. These forces both point towards the center of the Earth, so they are not parallel to each other. The force on the near side of the Moon tends to twist the Moon against its rotation, while the force on the far side tends to increase the rotation. But the force on the near side is stronger, because it is nearer the Earth, so the net effect is to slow the Moon's rotation down. At the same time, the near-side force also has a small component pushing the Moon forward in its orbit, while the far-side force does the opposite. Again, the near-side force wins, and the net push is along the orbital motion. This forces the Moon further from the Earth.

forces got larger: with more time to act on any region of the Moon, the tidal forces were able to raise larger tides in the rocks, and this accelerated the slowing down of the Moon. Eventually, the Moon came to present the same face to the Earth at all times, so the bulge was able to align exactly with the tidal force, and now the Moon no longer loses rotation. We say that the Moon is now in **synchronous rotation** with its orbit, since it rotates on its axis once each orbit.

The Moon's tidal effect on the Earth similarly tends to decrease the rotation rate of the Earth, but the Moon's effect is much weaker on the more massive Earth, so it has not yet brought the planet into synchronism with its orbit. A billion years ago, the day was only about 18 hours long.

The loss of spin by the Moon and the Earth has had another effect on the two: it has driven them further apart. The same tidal forces that have aligned the Moon's figure with the direction towards the Earth have also tended to give the center of the Moon a slight push in the same direction as it is orbiting the Earth, with the result that it has been flung away from the Earth. (You can see this if you study Figure 5.5 carefully.) The radius of the Moon's orbit has grown, and its orbital period increased.

In addition to the effects of the Earth on the Moon's orbit, there are small effects due to the Sun and the other planets. I have mentioned above how the plane of the Moon's orbit rotates because of the Sun. So too does the location of the perigee of its orbit, from the same cause. These effects are similar to those produced by Jupiter in the orbits of the other planets, which we shall discuss below.

►The increase in the radius of the Moon's orbit can also be understood as a consequence of the conservation of angular momentum, which we met in Chapter 4. The loss of spin by the two bodies increases the angular momentum of the orbit, forcing them apart.

►The *perigee* is the point of closest approach to the Earth. We saw in Chapter 4 that the nearest approach of a planet to the Sun is its *perihelion*, the suffix “-helion” referring to the Sun. The Moon or an Earth satellite has a *perigee*, “-gee” being a modification of “geo”, referring to the Earth.

In this section: tidal phenomena can be seen everywhere in astronomy. Mercury is tidally locked to its orbit, Jupiter's moon Io is heated so much by tides that it has volcanos, the asteroids are remnants of a failed attempt to form a planet too close to the tidal influence of Jupiter, and galaxies – whole systems of stars – crash together and disrupt one another tidally.

▷ The heat that volcanism on Io requires is far in excess of what can be being liberated by natural radioactivity in Io's interior. Radioactivity is thought to be the ultimate source of the heat that drives the Earth's volcanic and tectonic activity. All Earth rocks contain trace amounts of radioactivity, but when added up over the volume of the Earth's interior, the source of heat is enough to keep the interior molten. Volcanos burst out when hot molten rock manages to puncture the crust of the Earth at a weak point. Radioactivity does not force the molten rock out; it simply provides the heat that keeps it liquid.

Tides elsewhere in astronomy

Whenever two extended bodies are sufficiently near to one another, one can expect tidal forces to operate. Mercury is so close to the Sun that the tidal force of the Sun across it is nearly three times as large as those the Earth experiences from the Moon. Mercury orbits the Sun once every 88 days, and it turns on its axis once every 58.6 days, exactly 2/3 of its orbital period. This ratio isn't an accident: it is due to the tidal effects of the Sun.

Let us look at this another way, from the point of view of someone standing on Mercury. In two orbits Mercury spins three times. But this spin is with respect to the distant stars: during this time the person on Mercury has seen the stars go around three times. On the other hand, the Sun has also moved through Mercury's sky. Since Mercury has made two orbits of the Sun, the Sun has appeared to move twice through the sphere of stars, as seen from the ground, but in the opposite direction. So it has actually gone through Mercury's sky only once.

Mercury has a "tidally locked" day that lasts twice as long as its year.

(If it always presented the same face to the Sun, as the Moon does to the Earth, then its day would be infinitely long.) Calculations show that this arrangement can be stable if Mercury is not perfectly symmetrical about its rotation axis: Mercury's spin is probably not slowing down any more.

A spectacular example of the effects of tides is Io, the nearest to Jupiter of the four moons discovered by Galileo with his first telescope. When the Voyager 1 spacecraft flew by Jupiter in 1979 it observed no less than *eight* volcanic eruptions (see the image on page 39). These eruptions and the smooth, craterless surface of Io suggest that this volcanism has been going on at a steady rate for a very long time.

Io is heated by friction caused by the tidal forces of Jupiter on its moon. These are 250 times as strong as the Earth's forces on the Moon, chiefly because Jupiter is 300 times as massive as the Earth. Io is tidally locked to Jupiter, presenting the same face to it all the time. But it rocks back and forth about this position because of the tidal effects of Jupiter's other large moons, Europa and Ganymede. These two moons have orbital periods that are tidally locked to Io's: mutual gravitational forces between the moons have arranged that Io's period is half of Europa's and Europa's is half of Ganymede's. The regular tidal "bumping" of Io by these moons has built up a significant wobble, and the distortion of Io's tidal bulge during its wobble generates heat through friction inside Io. The distortion is not a small effect: parts of Io's surface can go up and down by as much as 100 m. It is not surprising that such large motions can lead to volcanism on this moon.

The asteroid belt, a system of large planetesimals orbiting the Sun between Mars and Jupiter, looks like the remnants of the formation of a planet that was stopped prematurely. The most likely cause is Jupiter's and the Sun's tidal forces: the weak binding forces holding together a pair of planetesimals was no match for the tidal forces. We shall discuss how this happens in more detail in Chapter 13.

Outside the Solar System tidal effects are also common. Most stars seem to be in binary systems, in which two or more stars orbit one another. Sometimes they can be quite close, closer than the Earth is to the Sun, or even in contact, so that their outer surfaces actually touch. In such situations, the shapes of the stars can become very distorted, and gas may even flow from one star to the other, sometimes with spectacular results. We will discuss these sorts of stars in Chapter 13.

Galaxies, too, can exhibit tidal effects. A galaxy is a collection of anywhere from 10^9 to 10^{12} stars, bound together by their mutual gravitational attractions. When two such galaxies get too close, the tidal forces of one can strip stars away from the

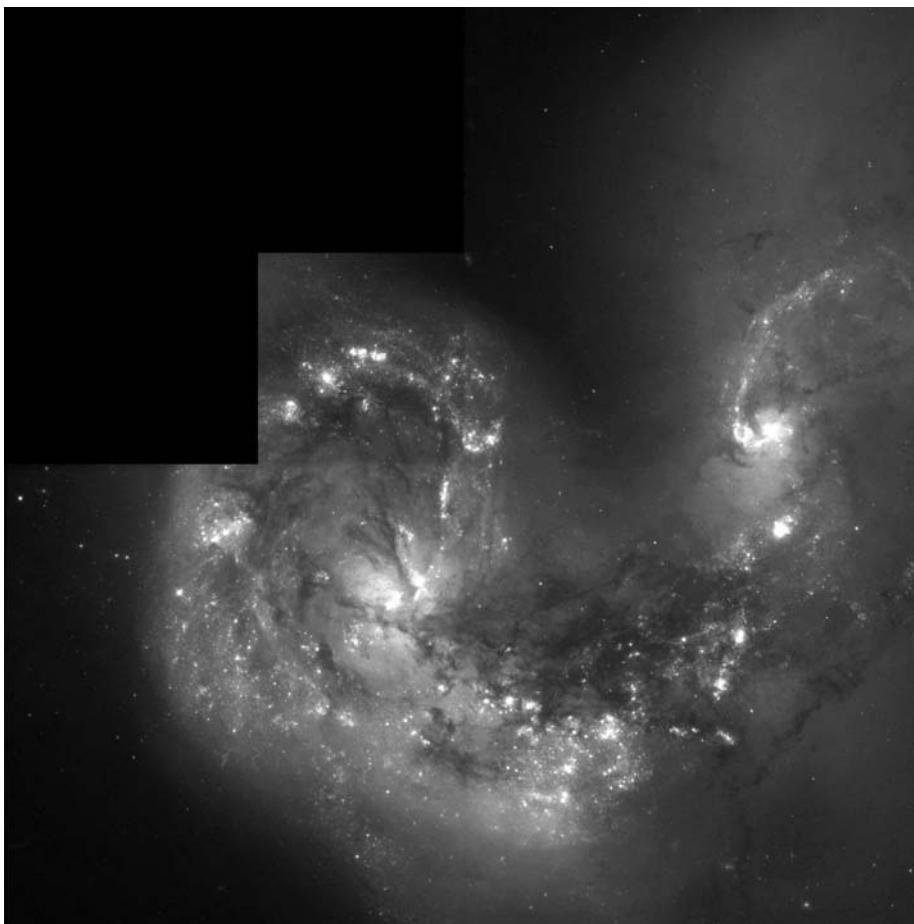


Figure 5.6. A photo-mosaic of the Antennae galaxies, taken by the Hubble Space Telescope (HST). These are two galaxies which are undergoing a collision. The non-uniformity of the gravitational accelerations of each galaxy on the other disrupt the normal orbital motion of the stars. The galaxies will eventually merge into a galaxy with a much smoother appearance. This is a snapshot of a collision that will take hundreds of millions of years to complete. The Milky Way and the great galaxy in Andromeda, our Galaxy's closest large neighbor, are similarly falling towards one another, and may collide like this in a billion years or so. (Courtesy NASA and its National Space Science Data Center (NSSDC).)

other, producing chaotic streams of stars. Pairs or groups of such interacting galaxies are a common sight in photographs taken by the biggest telescopes (Figure 5.6).

Our own galaxy, the Milky Way, may well now be showing the after-effects of such a tidal encounter. The Magellanic Clouds, bright patches of stars well away from the Milky Way in the sky visible from the Southern Hemisphere, are now known to be the brightest parts of a whole stream of stars extending right down to the Milky Way. The origin of this Magellanic Stream is still a matter of debate among astrophysicists, but one idea is that it may have been torn from the Milky Way by the tidal forces of another galaxy, or vice versa.

In fact, astronomers have discovered a region in the Milky Way, the other side of the center of the Galaxy from the Sun's location, where there is a large group of stars all traveling together with a different speed from most other stars. These stars may be the remnants of a small galaxy that is currently being torn apart and swallowed by the Milky Way. This may have happened many times in the history of the Milky Way. When such clumps get absorbed by the Milky Way, they go into orbits that retain a "memory" of how they fell in; they do not randomize their motions rapidly, because individual stars do not often come close enough together to deflect each other from their orbits. European astronomers are preparing a space mission called GAIA that could measure the speeds of stars all over the Milky Way

>The GAIA mission is one of the most ambitious space missions yet designed. For its wide range of scientific goals, see its website, <http://www.estec.esa.nl/spdwww/future/html/gaia.htm>.

so accurately that it would be able to identify such fossil “streams” of stars in the Milky Way, and thereby open a window into the past history of our own Galaxy.

Jupiter gives Mercury’s story another twist

In this section: the effect of the tidal forces of Jupiter on Mercury’s orbit is to push the ellipse around the Sun. This is called the precession of the orbit.

Nineteenth century astronomical observations of the motions of the planets became so precise that it was easy to see that the orbits of planets were not the perfect ellipses that one would expect if the Sun were the only gravitating body: the effects of the gravitational pull of the other planets caused slight but measurable deviations from ellipses. Many of these effects can be thought of as tidal effects.

Consider, for example, the effect that Jupiter has on Mercury’s orbit about the Sun. Mercury will go around the Sun several times while Jupiter changes its position only slightly. We have seen in Chapter 4 that the Sun executes a small orbit because of Jupiter’s gravitational pull on it. Mercury will follow the Sun as it does this, keeping the Sun at the focus of its elliptical orbit. So Mercury’s orbit will not remain a perfect ellipse relative to the stars: it is best described as an ellipse that changes its location gradually, as the Sun moves. The direction of the major axis of the ellipse does not change during this motion: the ellipse keeps its orientation.

All this is because the Mercury–Sun system falls freely in Jupiter’s gravitational field. But Jupiter’s tidal forces will have a further effect on the ellipse of Mercury’s orbit. We can see what to expect by thinking of Mercury’s orbit, not as empty space, but as a line along which Mercury’s mass is spread out. This is an acceptable approximation because Mercury orbits so much faster than Jupiter: it executes almost 50 orbits during one of Jupiter’s, so the mean gravitational effect of Jupiter is indeed spread out along Mercury’s orbit.

Now, when Jupiter is near one of the “bulges” in the ellipse of Mercury’s orbit, it will tend to pull that bulge toward it, just as the Earth tries to align the Moon’s bulge in Figure 5.5 on page 45. But Jupiter moves, while the direction of the bulge stays fixed in space. When Jupiter is approaching the bulge from behind it, it will tend to pull the bulge backwards. After it passes the bulge, it will tend to pull the bulge forward. But the first situation will last a little less time than the second: when approaching the bulge, the fact that Jupiter pulls it towards itself makes it reach the bulge more quickly than if it had not pulled the bulge, while the opposite happens after Jupiter passes the bulge.

The net effect, therefore, is to give the bulge a net pull in the direction of Jupiter’s orbit. The perihelion of Mercury’s orbit – its point of closest approach to the Sun – will move forward, i.e. in the same sense as Jupiter is moving. We say that Mercury’s ellipse *precesses* forwards. All the planets have similar effects on one another’s orbits. In the nineteenth century, mathematical physicists developed powerful approximation methods to calculate these precession effects; they needed them, because they did not have electronic computers! Much of modern mathematics has its roots in these calculations.

Triumph of Newtonian gravity: the prediction of Neptune

In this section: Neptune was discovered when it was shown that small unexplained motions of Uranus could be explained as tidal perturbations by an unseen planet.

Nineteenth century mathematicians found that they had to take into account all these small perturbations on the orbits in order to reconcile observations with Newtonian theory, and thereby to discover if Newtonian gravity was really an accurate description of gravity. Without the aid of electronic computers, the calculations involved were mammoth, and the mathematical techniques these scientists invented to simplify their job founded the modern branch of mathematics known as perturbation theory. The triumph of their calculational proficiency, and of Newtonian gravity, was the prediction by John C Adams (1819–1892) and Urbain Le Verrier (1811–1877) that certain perturbations of the orbit of Uranus could be explained

if there were another planet, further away from the Sun. The planet was indeed discovered near the predicted location in 1846, and named Neptune. Newtonian gravity seemed unassailable.

Tiny flaw of Newtonian gravity: Mercury's perihelion motion

The very calculations that gave Newtonian gravity its triumph also brought about its one failure in planetary theory: its inability to account for the full precession of Mercury's perihelion. Observations showed that Mercury's orbit precesses by about 574 arcseconds per century, about 0.16 degrees. Le Verrier, who had predicted Neptune, calculated in the 1850s that the effects of all the planets on Mercury could not account for the whole precession. By the 1880s, astronomers knew how to explain only 531 arcseconds per century, leaving 43 arcseconds per century unexplained.

Naturally, scientists tried the same route as for Neptune: postulate an extra planet or other sort of matter. But none was discovered. The problem became so serious that a modification of Newtonian gravity was proposed, to change the exponent 2 in the inverse-square law to something slightly larger than 2. Readers who play with the orbit program may experiment with other exponents, and should observe that orbits precess forwards if the exponent is increased a little, and backwards if it is decreased. Since the amount of precession was slight, the required change in the exponent was small; but even this turned out to be inconsistent with better and better observations of the Moon's orbit. Not until Einstein's theory of general relativity appeared was there a satisfactory explanation of this precession. We will return to this story in Chapter 18.

In this section: the tidal effects of all the planets together cannot account for the total precession of the orbit of Mercury. A tiny residual amount went unexplained until Einstein showed that his theory of general relativity predicted the effect exactly.

Interplanetary travel: the cosmic roller-coaster

Some of the most exciting moments in the exploration of space in the last thirty years have been provided by a succession of unmanned spacecraft that have explored more and more remote reaches of the Solar System. The early Moon-orbiters, scouts for later Moon landers, were succeeded by spacecraft that visited Mercury Venus, Mars, Jupiter, Saturn, Uranus, Neptune, various comets, and the Sun itself.

But to explore the Solar System in this way requires stronger and stronger rockets, much stronger than are required simply to get a spacecraft away from the Earth's gravitational pull. In order to do the most with the rockets available to them, planetary scientists have used a remarkable trick, called the **gravitational slingshot**: they have used the gravitational pull of another planet, such as Jupiter, to give their spacecraft an extra kick in the direction they want it to go. In this chapter we will try to understand how this works, not only for getting spacecraft into the outer parts of the Solar System, but also for getting them very close to the Sun.

Getting away from the Earth

We remarked in Chapter 4 that the *escape speed* from the Earth is 11.2 km s^{-1} , which is $\sqrt{2} = 1.414$ times the orbital speed at the Earth's surface. We shall prove this in Investigation 6.1 on page 53, which readers should read in connection with the next section. We will see there that, when launched with the escape speed, a spacecraft will just barely get away: if the Earth were the only gravitating body around, it would coast away at an ever-decreasing speed that would tend towards zero as it got far away. In the context of the Solar System, "far away" is still relatively near to the Earth. A spacecraft launched with the speed of 11.2 km s^{-1} in any direction from the Earth would soon find itself roughly stationary with respect to the Earth, i.e. orbiting the Sun with the same speed and therefore in roughly the same orbit as the Earth itself.

Getting away from the Earth to another planet therefore must require a launch with a speed greater than 11.2 km s^{-1} , but the result of such a launch will depend on the direction the spacecraft goes, relative to the Earth's motion around the Sun. If it is shot out in the forward direction, then its excess speed will add to the Earth's own orbital speed, and the result will be an orbit that carries the spacecraft farther from the Sun, in an orbit with a perihelion of 1 AU. This orbit will take the spacecraft outwards to other planets. If the spacecraft is shot in the backward direction, its excess speed will *subtract* from the Earth's speed, resulting in an orbit that falls in closer to the Sun.

Let us imagine an extreme case: trying to get a spacecraft completely out of the Solar System. Since the Earth's orbit is roughly circular, the escape speed from the Sun is 1.414 times the Earth's orbital speed of 29.8 km s^{-1} (see Table 4.2 on page 28), which makes 42.1 km s^{-1} . This is the speed the spacecraft must have, relative to the Sun. When we launch the spacecraft, it already has the same speed as the Earth

In this chapter: mastering interplanetary navigation has opened up the planets to exploration in the last 50 years. The discoveries have been astonishing. The motion of spacecraft teach us much about mechanics: about energy and the way it changes, about momentum and angular momentum, and deepest of all about the role that *invariance* plays in modern physics.

In this section: we learn how to get enough speed to reach other planets, and shows that it is harder to get to the Sun than to escape from the Solar System.

while it sits on the launch pad. We can use this fact to take maximum advantage of the Earth's orbital speed, by shooting the spacecraft directly forward in the Earth's orbit, so that we only need to "top up" the speed by another 0.414 times the Earth's speed, or 12.3 km s^{-1} . This is, of course, the speed *after escaping from the Earth's gravity*. If we send the spacecraft out in any other direction, the Earth's speed will not contribute so much to its final speed in that direction, so it will require more of a boost to get it away from the Solar System.

To get the launch speed from this we have to add to the final speed the Earth's escape speed, 11.2 km s^{-1} . This gives a minimum launch speed that is more than twice as large as the escape speed from the Earth itself. This is rather large, and would require a powerful rocket and a great deal of fuel. We will see that it is cheaper to use the rocket to get as far as Jupiter and then to use the slingshot mechanism to get further.

Surprisingly, it is even harder to send a spacecraft very close to the Sun than to escape from the Sun altogether: for this we must insure that after it escapes from the Earth the spacecraft stops nearly dead in its orbit *relative to the Sun*, so that it can fall straight in towards it. This means that it must be shot out in a backwards direction so that its excess velocity relative to the Earth after escaping from the Earth is nearly equal to the Earth's orbital speed around the Sun, 29.8 km s^{-1} . When added to the escape speed, this requires more than three times the Earth's escape speed, so it follows that it would require a much bigger rocket to reach the Sun than that required to get away from the Solar System entirely. Here, too, we shall see that it is better to send such a spacecraft to *Jupiter* first, and let Jupiter direct it towards the Sun!

Plain old momentum, and how rockets use it

In this section: we learn about ordinary momentum and use it to explain how rockets work.

We have met, in this chapter as well as in Chapter 4, the idea of the conservation of angular momentum, which governs orbits around the Sun. We shall now add to this the law of conservation of (ordinary) **momentum**, which will help us understand how rockets move around in the Solar System.

The momentum of a body, say a rocket, is defined as the product of its mass m and its velocity.

$$\text{momentum} = \text{mass} \times \text{velocity}. \quad (6.1)$$

It is important to distinguish between the rocket's speed and its velocity, because the velocity depends on the direction the rocket is going in. One of the deep laws of physics is:

the total momentum of a collection of bodies is constant in time if there are no forces acting on it from outside.

This law is called the law of *conservation of momentum*. This can help us understand how rockets propel themselves. Rockets carry their own fuel. They burn it at a controlled rate and expel the exhaust gases out the back. This accelerates the rocket forwards. How does this happen: how does having a hole at the back help the rocket move forward?

The gases that come out of the rocket nozzle have a small mass compared to that of the rocket, but they have a large speed. So they carry a lot of momentum. There are no forces acting on the rocket from outside (let's forget the small effect of gravity for the purposes of this discussion), so the total momentum of the rocket plus gases is constant. Therefore, the momentum carried away by the gases is lost by the rocket. But this momentum is directed backwards, so it is negative: the velocity of the exhaust gases is a negative number. The rocket loses this negative number from

Investigation 6.1. Escaping – you can get away from it all if you have enough energy

Energy holds the key to deciding whether a satellite will be able to escape from the Sun or not. Let us look at how much energy an escaping orbit has.

We look at the detailed form of the definition of the total energy, from Equations 6.8 and 6.9:

$$E = \frac{1}{2}mv^2 - \frac{GmM_{\odot}}{r}. \quad (6.2)$$

A body has escaped from the Sun if it can get arbitrarily far away, i.e. if we can make r as large as we want. This means that the second term in this equation can be made as small as we like, so that eventually the body is coasting with a constant speed v_{far} given by

$$E = \frac{1}{2}v_{\text{far}}^2.$$

A body just barely escapes if its final speed is zero, which means that the total energy on its trajectory is zero: *the trajectory that only just escapes has zero total energy*. It turns out that this is true as well if we turn the sentence around: if a trajectory has total energy zero, then it is the path of a body that will get arbitrarily far from the Sun, but whose speed goes to zero as the distance gets larger.

Now, suppose the body starts out at a distance R from the Sun. In order to follow a trajectory of zero energy, it must have a speed v_{escape} given by setting E to zero in Equation 6.2:

$$\frac{1}{2}mv_{\text{escape}}^2 = \frac{GmM_{\odot}}{R} \Rightarrow v_{\text{escape}} = \left(\frac{2GM_{\odot}}{R} \right)^{1/2}. \quad (6.3)$$

Exercise 6.1.1: Escaping from anywhere

Calculate the escape speed from the Solar System for a satellite starting at the average distance from the Sun of each of the planets listed in Table 4.2 on page 28. In each case, find the ratio of this speed to the average speed of the planet (column 5 of the table).

its momentum, and this means it actually *increases* its momentum. The equation looks something like this:

momentum gained by rocket = -momentum carried away by gases.

The minus sign cancels the minus sign in the momentum of the gases and leads to an increase in the momentum of the rocket. This is called the rocket equation.

There is a deep relation between this equation and Newton's laws of motion. There has to be: we should equally well be able to explain the acceleration of the rocket by the fact that the exhaust gases exert a force on it, propelling it forwards. Then we would use $F = ma$ to calculate the acceleration a . However, here we have to be careful: when Newton wrote down this equation he assumed that bodies would keep the same mass as they accelerate. But the rocket does not: its mass is always changing. We can see that Newton's second law has a slightly different form in this case from the following argument.

The acceleration a is the change in velocity divided by the time-interval during which the velocity changes. If we multiply both sides of the equation $F = ma$ by this time-interval, we get something that reads:

$$\text{Force} \times \text{time-interval} = \text{mass} \times \text{change in velocity}.$$

For a body with constant mass, the right-hand side is the same as the change in the product of mass and the velocity, or the momentum. So for such a body, an equivalent expression is

$$\text{Force} \times \text{time-interval} = \text{change in momentum}.$$

This is the escape speed from the Sun. For other bodies, we just replace M_{\odot} with the mass of the body from which we are escaping. We quoted this result in Chapter 4.

It is interesting to compare the escaping orbit with a circular one. Since the acceleration of gravity by the Sun at this distance is $g = GM_{\odot}/R^2$, it follows from Equation 3.1 on page 19 that the circular orbital speed is

$$v_{\text{circ}} = \left(\frac{GM_{\odot}}{R} \right)^{1/2}, \quad (6.4)$$

which means that

$$v_{\text{escape}} = \sqrt{2}v_{\text{circ}}. \quad (6.5)$$

The total energy of a circular orbit is simple, as well:

$$E_{\text{circ}} = -\frac{GmM_{\odot}}{2R} = \frac{1}{2}V. \quad (6.6)$$

Put another way, the energy equation for a circular orbit implies

$$E = K + V = \frac{1}{2}V \Rightarrow 2K + V = 0. \quad (6.7)$$

Do not be worried by the fact that the total energy is negative. The only thing that is ever measurable is the change in the total energy from place to place or from orbit to orbit. If we were to add some huge constant energy to all energies to turn them into positive quantities, we would still have the same differences between energies, and the same physics.

These two equations are not equivalent for a body whose mass changes during the time-interval; that is, their right-hand sides are not the same. In an extreme case, a rocket could change its momentum by losing mass without changing its speed (say, by just pushing one of the astronauts gently out the door!), so the right-hand side of the second equation would be non-zero but that of the first would vanish. They can't both equal the force times the time-interval. So which one is right?

Newton's third law (Chapter 2) tells us the answer. Let us think about the rocket again. The exhaust gases exert a certain force F on the rocket, during whatever time-interval we want to consider. By the third law, the rocket exerts the equal and opposite force, $-F$, on the gases. So the left-hand sides of the two equations above are equal and opposite for the rocket and the gases. If the second equation is the correct one, then we will conclude that the change in the momentum of the gases is equal and opposite to the change in the momentum of the rocket, which is another way of saying that the total momentum is constant: momentum is conserved. The first equation is not equivalent to this, and so would not give conservation of momentum. Therefore, the second equation above is the correct version of Newton's second law to use when the mass of a body is changing. We will come back to this in Chapter 15 when we discuss the acceleration of bodies that go close to the speed of light.

Energy, and how planets never lose it

In this section: we define the total energy of an orbiting planet and show that it is constant.

Our discussion of orbits in the Solar System will be simpler if we first verify a third conservation law, the law of **conservation of energy** during the motion of a planet or spacecraft around the Sun. This is one of the most remarkable and profound ideas in all of physics.

In everyday language, "energy" is a measure of activity, or at least readiness for activity; and after a long period of activity we usually use up our energy and get tired. In physics, **energy** has a more precise definition, but one that is not unrelated to the everyday one. Here we will only discuss the physicists' definition of the energy of a planet; in later chapters we will begin to see its relation to our personal version of energy.

A planet of mass m moving in orbit around the Sun, has two kinds of energy. It has energy associated with its total speed v , called its *kinetic energy*,

$$\text{kinetic energy } K = \frac{1}{2}mv^2, \quad (6.8)$$

and it has another kind of energy by virtue of its distance r from the Sun, called its *gravitational potential energy*,

$$\text{potential energy } V = -\frac{GmM_{\odot}}{r}. \quad (6.9)$$

The symbol M_{\odot} is the symbol astronomers always use for the mass of the Sun, which is about 2×10^{30} kg. The standard (si) unit that scientists use for energy is the **joule**, abbreviated J. In Equation 6.8, the units on the right-hand side work out to be $\text{kg m}^2 \text{s}^{-2}$; one $\text{kg m}^2 \text{s}^{-2}$ is, by definition, equal to one joule.

A more familiar concept in everyday life is the related unit for **power**, which is defined to be the rate at which energy is used: energy per second. This unit is the **watt** (W), equal to 1 J s^{-1} . Thus, a 100 W light bulb consumes 100 J of electrical energy every second. We have not yet made the connection between the energies of a body's motion and other energies, like electrical. We will not do that here, but we will return to the subject and develop it further at several points in later chapters.

We shall not try to justify the precise forms of the definitions given above of kinetic and potential energy; rather we shall investigate them "experimentally", by

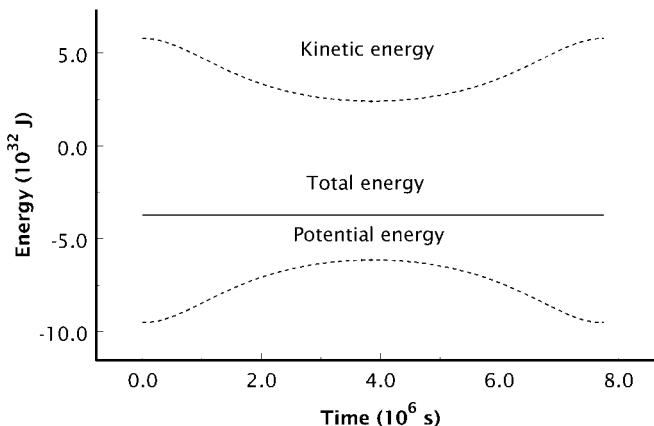


Figure 6.1. The law of conservation of energy is illustrated for Mercury's orbit by using data from the computer calculation of Mercury's orbit. Although the kinetic and potential energies, as defined in Equations 6.8 and 6.9, respectively, change a good deal during the planet's orbit, their sum remains constant.

using our computer program to show that the total energy, the sum of the two, is constant.

Before that, let us insure that we see that the individual definitions make sense. The kinetic energy increases as the speed increases, so it certainly measures how "energetic" the motion is. Moreover, bodies that are more massive also have more kinetic energy at any given speed. The potential energy at first looks strange, since it is negative. Far from the Sun it is essentially zero, because of its dependence on $1/r$. As r gets smaller, so the planet gets closer to the Sun, the potential energy gets larger in absolute value, and hence more negative. So the potential energy *decreases* as we get nearer the Sun. This energy is called "potential" in the sense that a planet far from the Sun has more potential to increase its speed and hence its kinetic energy by falling in towards the Sun than does a planet that is already close to the Sun, where the potential energy is less. We will return to why scientists use the word "potential" below, after defining the conservation law for energy.

Now we can formulate the law of conservation of energy: for a planet orbiting the Sun, the value of the total energy $E = K + V$ is constant everywhere along the orbit.

As in Chapter 4, we verify this conservation law by using the orbit program to calculate the values of the kinetic, potential, and total energies for the orbit of Mercury. These are displayed in Figure 6.1. It is clear that the total energy remains constant even as the kinetic and potential energies change. Readers are encouraged to verify this for any other orbits they may have calculated. We shall take this as a sufficient verification of the law, so that we can use it in our discussions of interplanetary travel below.

We can now understand the name "potential energy" a little better. Since the sum $K + V$ is constant, an orbit can go from a region in which K is small to one where K is large. But K cannot grow indefinitely large: the difference in V in the two places gives the change in K , so that V really does contain the potential increase in K along the orbit.

Conservation of energy has important uses when we try to understand the orbits of bodies around the Sun. The most important of these is to imply that there is a single *escape speed* that any body has to reach in order to get away from the Sun, regardless of the direction it takes. This speed depends only on how strong gravity is where the body starts out. If the body is a distance R from the Sun, then its escape

speed is

$$v_{\text{escape}} = \left(\frac{2GM_{\odot}}{R} \right)^{1/2}. \quad (6.10)$$

If it reaches this speed it will be able to move into interstellar space.

Getting to another planet

In this section: how to use conservation of energy to discover the limits on a space probe's motion among the planets.

Before we see how the slingshot works, we have to see how we can get to other planets in the first place. Here our orbit program should help us. Suppose, for example, we want to reach Mars. In principle all we would have to do is to run the program a few times with various values of the initial velocity at the Earth's position, to see what is the minimum initial speed relative to the Earth that will just get us to the planet.

In practice, this trial-and-error method would be painfully slow unless one had a very fast computer. Instead, we shall show in Investigation 6.2 on page 58 how to calculate the right speed from the law of conservation of energy and Kepler's first law. After that we can use the computer simply to verify that the answer we get really does work.

The best strategy for launching a spacecraft so that it just barely reaches one of the outer planets is to take as much advantage as possible of the velocity of the Earth and shoot the spacecraft forwards in the Earth's orbit. Since the spacecraft already has the Earth's velocity when it sits on the launch pad, this strategy means that we have only to supply the excess speed required beyond the Earth's speed. This keeps the launch cost of the spacecraft as small as possible. Such an orbit has a perihelion at the Earth and it reaches its maximum distance from the Sun at Mars' orbit.

Suppose we launch from the Earth, at a distance R_1 from the Sun, at a target planet a distance R_2 from the Sun. Let r denote the ratio R_2/R_1 ; then r is just the orbital radius of the target planet expressed in astronomical units (AU). We show in Investigation 6.2 on page 58 that, in order just to reach the outer planet, the spacecraft is required to have a speed relative to the *Sun* after escaping the Earth of

$$v = \left(\frac{2r}{r+1} \right)^{1/2} v_{\oplus}, \quad (6.11)$$

where v_{\oplus} is the Earth's orbital speed, 29.8 km s^{-1} . For a trip to Mars, where $r = 1.52$, we find that v must be 32.7 km s^{-1} (Exercise 6.2.2 on page 58). To reach Jupiter, we require 38.6 km s^{-1} .

These are speeds relative to the Sun after escaping the Earth. To get the launch speed in each case, we have to subtract the Earth's orbital speed of 29.8 km s^{-1} and to add the 11.2 km s^{-1} Earth escape speed. For Jupiter, this works out to a launch speed of exactly 20 km s^{-1} , which is significantly smaller than the 23.5 km s^{-1} that is required to escape the Solar System.

It is easy to use the program *Orbit* on the website to verify these numbers, and therefore to provide an "empirical" verification of the above formula. For example, I ran the orbit program for a trajectory that just reaches Jupiter, using an initial position of 1 AU from the Sun, an initial velocity of 38.6 km s^{-1} parallel to the Earth's orbit, and a time-step of 8000 s. The calculated orbit reached a maximum distance from the Sun of 5.24 AU, close enough for our purposes to Jupiter's actual distance of 5.2 AU. When it reached Jupiter its speed was 7.36 km s^{-1} , which is again close enough to the value of 7.42 km s^{-1} predicted by Equation 6.13 or Equation 6.15 on page 58. The spacecraft took 2.9 years to reach Jupiter's orbit. Naturally, to encounter Jupiter, the launch must be timed correctly, so that Jupiter is in the right position in its orbit when the spacecraft arrives. The orbit program can give the

information required for this as well. The reader is encouraged to try the program for some other planets, to see how long it would take to reach them.

As just noted, reaching the orbit of another planet is fruitless unless the planet is there to encounter the spacecraft at the right time. This means that there has to be a favorable disposition of the planets to allow a launch, and this may happen only once a year for the direct trajectories we have discussed so far, and less frequently for the slingshot effect. If one is unable to use the best “launch window”, then reaching the target planet will require a greater launch speed. The penalty can be severe. For example, if one tried to reach Jupiter on a trajectory that went purely radially outwards from the Earth’s orbit, it turns out that this would require a speed relative to the Earth after escape of 48.2 km s^{-1} . (You could try to verify this using the orbit program as well. If you do, bear in mind that the program requires the speed of the spacecraft relative to the *Sun*, which for this case is 37.9 km s^{-1} directly away from the Sun.) Thus, it is important for interplanetary missions that the launches take place on schedule!

The principle of the slingshot

Now that we know how to get a spacecraft to Jupiter, we can start thinking about how Jupiter can give it a further push to go somewhere else. The basic idea of this gravitational slingshot is that Jupiter (or another planet) supplies the extra energy that we could not get from our rocket engines. The energy comes from Jupiter’s motion, but Jupiter is so massive and has so much energy of motion (its kinetic energy) that these encounters make no significant change in Jupiter’s own orbital motion.

However, underneath this simple statement — that Jupiter will give our spacecraft a push — is a subtlety that we encounter immediately if we try to understand how it works. Is it really possible for Jupiter to give a spacecraft a push at all? After all, our orbit program shows no such effect: if we send a spacecraft on a trajectory around the Sun, and we follow its orbit as it falls towards the Sun and then comes back out, we always find that it returns to the same place as we started it with exactly the same speed. We made a point of showing this, because it is required by the fact that orbits in Newtonian gravity are closed. So the Sun doesn’t give our spacecraft any extra push: it doesn’t have any more speed after its encounter with the Sun than it had before. The same would be true if we did an orbit calculation for a spacecraft that does not stay in orbit about the Sun, but instead falls towards the Sun with a large initial speed; when the spacecraft returns to the radial distance from which it started, it has the same speed as it started with, although in a different direction. This is just a consequence of the conservation of energy: since its potential energy V depends only on the distance of the spacecraft from the Sun, it follows that its kinetic energy, and therefore its speed, also depends only on the distance from the Sun. So if the Sun doesn’t give anything an extra push, then how will Jupiter be able to do it?

The resolution of this apparent contradiction is to remember that all speeds are meaningful only when referred to some standard of rest. In the orbit program, we take the Sun as our standard of rest, and so the correct statement is that a spacecraft that encounters the Sun will have the same speed *relative to the Sun* after the encounter as it had before. The same will be true of encounters with Jupiter: it will have the same speed *relative to Jupiter* after the encounter as it had before. But for the spacecraft traveling through the Solar System, the important speed is not its speed relative to Jupiter, which is unchanged by the encounter, but its speed *relative to the Sun*, which can indeed change.

Let us look at a simple example of using Jupiter as a slingshot. Suppose a space-

In this section: we see how Jupiter can accelerate a space probe even though the energy of the space probe’s orbit around Jupiter is conserved.

Investigation 6.2. The reach of an orbit

The range of distances that a spacecraft can explore is governed by its energy and by Kepler's first law, which we met in Chapter 4. In this investigation, we will use these laws to get the speed we have to give a spacecraft to move it from one place to another.

We consider the energy of an orbit that is required to go from a minimum distance R_1 to a maximum distance R_2 from the Sun. Suppose it has a speed v_1 at R_1 , its perihelion. As we know, it follows an ellipse, and arrives with a smaller speed, v_2 , at R_2 , which we call its **aphelion**, its furthest distance from the Sun. The equation of energy conservation is one equation that relates these four quantities. For any given orbit, the quantity E in Equation 6.2 on page 53 must be the same wherever it is calculated. This implies

$$\frac{1}{2}mv_1^2 - \frac{GM_{\odot}}{r_1} = \frac{1}{2}mv_2^2 - \frac{GM_{\odot}}{r_2}. \quad (6.12)$$

We can get another relation from Kepler's first law. Recall that this says that the area swept out in any fixed time Δt by the line from the Sun to the planet is the same anywhere along the orbit. (This is also called the law of conservation of angular momentum.) Although this could be difficult to work out at a general position along an elliptical orbit, it is not hard at the perihelion and aphelion, where the velocity is momentarily perpendicular to the direction to the Sun. If we consider a small time Δt just as the planet is passing perihelion, the planet will move a distance $v_1\Delta t$ in this time, and the small triangle in Kepler's law will have equal sides of length R_1 . (At other points of the orbit, these sides would not be equal and the calculation of the triangle's area would be more difficult.) The area of a triangle is one-half its base times its height. This triangle's base is $v_1\Delta t$, and its height is R_1 , to an excellent approximation. So it has area

$$\text{area at perihelion} = \frac{1}{2}R_1 v_1 \Delta t$$

A similar calculation at aphelion gives

$$\text{area at aphelion} = \frac{1}{2}R_2 v_2 \Delta t.$$

Kepler's first law says these are equal, so we have

$$\frac{1}{2}R_1 v_1 \Delta t = \frac{1}{2}R_2 v_2 \Delta t.$$

Cancelling out the factors of $\frac{1}{2}$ and Δt , we find a second and very simple relation among the distances and speeds at perihelion and aphelion:

$$R_1 v_1 = R_2 v_2. \quad (6.13)$$

If we solve Equation 6.13 for v_2 and substitute the result into Equation 6.12, we get an equation that we can solve for R_2 in terms of R_1 and v_1 : in other words, we can predict the aphelion of an orbit if we are given its perihelion distance and speed. After multiplying by $2/m$, this equation can be put into the form

$$v_1^2 - \frac{2GM_{\odot}}{R_1} = \frac{R_1^2 v_1^2}{R_2^2} - \frac{2GM_{\odot}}{R_2}.$$

Exercise 6.2.1: Solving the quadratic equation

The general solution of the quadratic equation $ax^2 + bx + c = 0$ for x is

$$x = -\frac{b}{2a} \pm \frac{1}{2a} \left(b^2 - 4ac \right)^{1/2}, \quad (6.17)$$

where the \pm sign indicates that there are two solutions, found by taking either sign in the expression. Apply this formula to solve the quadratic equation above for R_2 . Show that the two roots are R_1 and the root given by Equation 6.14.

Exercise 6.2.2: Getting from the Earth to other planets

Use Equation 6.16 to calculate the speed needed to go from the Earth's orbit to the orbits of Mars, Jupiter, and Saturn. The derivation of this formula actually did not need to assume that $r > 1$, so use it for Venus, too.

This can be simplified by introducing the symbol L_1 to denote the ratio

$$L_1 = \frac{GM_{\odot}}{v_1^2},$$

which is nothing more than the radius of a *circular* orbit about the Sun that has orbital speed v_1 . (This will lie somewhere between R_1 and R_2 .) The equation for R_2 now becomes, after dividing by v_1^2 , multiplying by R_2^2 , and arranging terms,

$$\left(\frac{2L_1}{R_1} - 1 \right) R_2^2 - 2L_1 R_2 + R_1^2 = 0.$$

This is a **quadratic equation** for R_2 . It would be easy to use the general solution for such an equation (Exercise 6.2.1), but we can do something even simpler by observing that one solution of this equation must be R_1 itself: R_1 is a place where the velocity is perpendicular to the radius, so both of our original Equations 6.12 and 6.13 apply. This means that $R_2 - R_1$ is a *factor* of the above equation, which can in fact be written

$$(R_2 - R_1) \left[\left(2 \frac{L_1}{R_1} - 1 \right) R_2 - R_1 \right] = 0.$$

The second factor provides the other solution, for the aphelion:

$$\left[\left(2 \frac{L_1}{R_1} - 1 \right) R_2 - R_1 \right] = 0,$$

which gives finally

$$R_2 = R_1 \left(\frac{2GM_{\odot}}{R_1 v_1^2} - 1 \right)^{-1}. \quad (6.14)$$

This is the desired expression for the aphelion distance in terms of the perihelion distance and speed. We can put this back into Equation 6.13 to solve for v_2 , the aphelion speed:

$$v_2 = \left(\frac{2GM_{\odot}}{R_1 v_1^2} - 1 \right) v_1. \quad (6.15)$$

We can now ask the question we had in mind from the beginning, which is to find the speed v_1 that we need to give to a space probe to get it from the Earth's orbit at R_1 to Jupiter's at R_2 , provided it is fired straight ahead along the Earth's orbit. It is convenient in Equation 6.14 to replace GM_{\odot}/R_1 by v_{\oplus}^2 , the square of the Earth's (circular) orbital speed. Then solving for v_1 gives

$$v_1^2 = \frac{2r}{r+1} v_{\oplus}^2, \quad \text{where } r = R_2/R_1. \quad (6.16)$$

This is equivalent to Equation 6.11 on page 56.

craft approaches Jupiter along Jupiter's own orbit, catching up with it from behind. Jupiter's orbital speed is 13.1 km s^{-1} , and let us assume for this illustration that the spacecraft is going at 15.1 km s^{-1} relative to the Sun. Then it is approaching Jupiter at a relative speed of 2 km s^{-1} . Again to make the illustration simple, let us assume that the encounter turns the spacecraft completely around, so that afterwards it leaves Jupiter going back toward where it came from. It will leave Jupiter with same the relative speed of 2 km s^{-1} , but now this is directed backwards along Jupiter's orbit, so that the resulting speed relative to the Sun is only 11.1 km s^{-1} . The encounter has slowed the spacecraft down *relative to the Sun*.

This is the sort of trajectory one would look for in order to send a spacecraft close to the Sun. Alternatively, we could have arranged for the spacecraft to approach Jupiter from the other direction, and the result would have been to speed it up and send it further out in the Solar System.

So have we partly lost conservation of energy? Is energy conserved when we measure it relative to Jupiter but not relative to the Sun? No; if that were the case then the law would not be a law at all. If we go back to measuring speeds relative to the Sun, then we have to take into account all the energies, both that of the spacecraft and of Jupiter. If the encounter speeds up the spacecraft, then it must slow down Jupiter. But because Jupiter's mass is so large, the change in its speed is too small to notice. Conservation of energy is fine, but the kinetic energy of a planet is so large that it is an essentially infinite reservoir on which we can draw for our planetary explorations.

Using Jupiter to reach the outer planets

Are the numbers we have quoted earlier realistic? If not, how effective could Jupiter be in a real situation? If we want to reach, say, Saturn from Jupiter's orbit, then we can use our previous formula to tell us the minimum speed we need to have when we leave Jupiter's orbit. Taking $r = 1.83$, which is the ratio of the orbital radii of Saturn and Jupiter, and using Jupiter's speed of 13.1 km s^{-1} in place of v_{\oplus} in Equation 6.11 on page 56, we find that we need a minimum speed relative to the Sun of 14.9 km s^{-1} to reach Saturn from Jupiter. This means we need to leave Jupiter with a speed of at least 1.8 km s^{-1} relative to it, in the forward direction in its orbit.

If we could get an encounter with Jupiter that turned the spacecraft entirely around, we therefore would need to have reached Jupiter's orbit at a point slightly *in front* of Jupiter with a speed 1.8 km s^{-1} *lower* than the speed of Jupiter in its orbit, so that the spacecraft effectively approaches Jupiter from the front. This is an orbital speed of 11.3 km s^{-1} . This is the *maximum* speed we could allow in Jupiter's orbit for the slingshot mechanism to work: a higher speed would mean a lower speed of approach between the spacecraft and Jupiter and consequently a smaller boost from Jupiter of the spacecraft's speed.

The actual speed of the spacecraft when it reaches Jupiter is, as we have seen above, about 7.4 km s^{-1} , directed along the orbit of course. This is considerably below the maximum allowable for the mechanism to work, which means that we have plenty of leeway for playing with such things as the trajectory of the orbit to Saturn.

In fact, this margin allows us the flexibility to cope with another effect that we have so far ignored: a real encounter does not usually turn the spacecraft around by 180° . The angle by which the incoming and outgoing directions of the spacecraft relative to Jupiter differ is determined by the spacecraft's speed approaching Jupiter and by how close it actually approaches Jupiter. For safety reasons, the spacecraft must be kept well away from the planet's surface. So we cannot expect to get the full boost from Jupiter that our simple arguments suggest.

Even given this limitation, we could in principle use the slingshot to boost us

In this section: we examine the details of using Jupiter to boost the speed of a space probe.

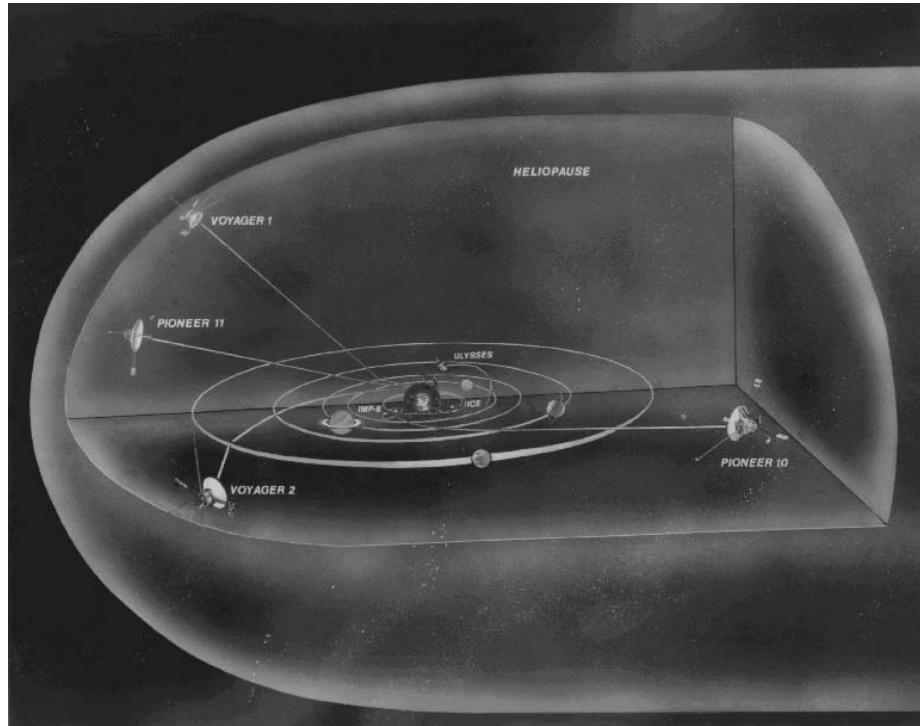


Figure 6.2. The trajectories of Pioneers 10 and 11 and Voyagers 1 and 2, showing encounters with planets that have sent all four spacecraft out of the Solar System.
(Courtesy JPL/NASA/Caltech.)

far beyond Saturn. Our spacecraft is closing with Jupiter at a speed of 5.7 km s^{-1} , so that (again, if it could be turned around and sent forward by Jupiter) its speed leaving Jupiter would be $5.7 + 13.1 = 18.8 \text{ km s}^{-1}$. The escape speed from the Sun from Jupiter's orbit is $13.1\sqrt{2} = 18.5 \text{ km s}^{-1}$, so a slingshot from Jupiter might propel the spacecraft out of the Solar System entirely. This is in fact what has happened to both Voyager spacecraft, although their trajectories are complicated by encounters with several planets (Figure 6.2).

Interestingly, the Earth itself can be used to provide a slingshot to propel a spacecraft out to Jupiter and beyond. In Figure 6.3 we illustrate the trajectory of the Galileo spacecraft, which is a Jupiter probe. Its orbit first fell towards the Sun from the Earth, was boosted by Venus, then encountered the Earth twice more, propelling it out to Jupiter.

Slinging towards the Sun

In this section: we consider other possibilities for the slingshot mechanism, such as reaching the Sun or using the inner planets.

Comets reach the inner Solar System by the slingshot mechanism, mainly using Jupiter.

What about using Jupiter to put a spacecraft near the Sun? Here we ask Jupiter to slow the spacecraft down. To reach within, say, 0.1 AU of the Sun, a spacecraft must have an orbital speed at Jupiter's orbit of no more than 2.5 km s^{-1} , which we can again obtain from Equation 6.11 on page 56 with $r = 0.1/5.2$ and v_{\oplus} replaced by Jupiter's speed of 13.1 km s^{-1} . This orbital speed represents a speed relative to Jupiter of 10.6 km s^{-1} in the backwards direction. Before its encounter with Jupiter the spacecraft had to have the same speed of approach relative to Jupiter.

This situation is more difficult to analyze, because here one would want to take advantage of the fact that the spacecraft's trajectory is not deflected through 180° by Jupiter. One can see roughly how to make it work by considering two extreme cases that both result in trajectories with a speed of 2.5 km s^{-1} , one in which the spacecraft is indeed turned completely around and one in which there is no encounter at all. In the first case we approach Jupiter from behind going faster than it, and in the

second we approach from the front. The first requires an initial spacecraft speed of $13.1 + 10.6 = 23.7 \text{ km s}^{-1}$ relative to the Sun, while the second requires the same 2.5 km s^{-1} initially as the craft will have finally.

Neither speed is available to us, the first because it requires too much energy and the second because it is smaller than our spacecraft will have when it reaches Jupiter. But somewhere in between these two extremes is a trajectory that approaches Jupiter from the side with a speed relative to the Sun more like 6 km s^{-1} , and which will leave Jupiter in exactly the backwards direction provided we arrange the angle of deflection of the orbit correctly. (This is determined by the distance of closest approach to Jupiter.) Given that the two extreme speeds are 23.7 and 2.5 km s^{-1} , it seems clear that the real case will be closer to the second case, where we approach Jupiter nearly head-on from in front. Only a relatively small deflection will be required to remove a few kilometers per second of speed to get the spacecraft down from 6 to 2.5 km s^{-1} . We will not do this calculation in any greater detail here.

Interested readers may wish to consider other ways of reaching the inner Solar System, such as using an encounter with Venus to reach Mercury. Another option would be to approach Jupiter partly from below the plane of its orbit; diving under it this way could produce an orbit that is out of the plane of the planetary orbits, and which would then pass over the poles of the Sun. All of these tricks have been used or proposed for interplanetary exploration.

Artificial spacecraft are not the only objects that experience the slingshot mechanism. Outside the orbit of Neptune, stretching over many hundreds of AU, is the **Kuiper Belt**, a zone full of planetesimals that never formed into planets. It is named after the Dutch astronomer Gerard Peter Kuiper (1905–1973). Astronomers have only recently discovered how abundant these asteroids are, and how various their sizes are. In fact, Pluto seems more aptly described as a giant asteroid from the Kuiper Belt rather than a planet; its properties are very different from the gas giants, but very similar to the asteroids. Sometimes asteroids from the Kuiper Belt reach the orbit of Jupiter or Saturn, either because they have elliptical orbits or because they have collided with other asteroids. They can be slung by one of these planets into an orbit that takes them much closer to the Sun, and then they present a danger to the Earth. The collision that is thought to have assisted in the extinction of the dinosaurs was probably with one of these objects.

Outside the Kuiper Belt is the **Oort Cloud**, home of the comets. This region, extending as much as 10^5 AU from the Sun, has been named after another Dutch astronomer, Jan H Oort (1900–1992). The comets are believed to resemble the building blocks out of which the asteroids and from them the planets were formed. At distances so far from the Sun that they cannot collide often enough to build planets, these objects remain museum pieces of the earliest stage of planetary formation. There is great interest among astronomers in studying comets to learn about this period. If it were not for Jupiter, and to some extent Saturn, we would not see any in the inner Solar System. Although comets appear to have very eccentric orbits,

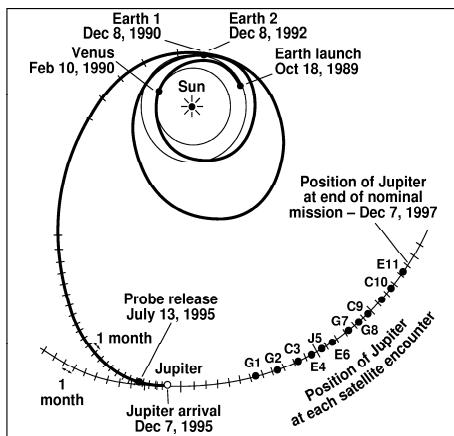


Figure 6.3. The trajectory of the Galileo spacecraft shows how the Earth itself can be used as a slingshot. (Courtesy JPL/NASA/Caltech.)

they rarely fall close to the Sun. Instead, when they reach the orbit of Jupiter they are slung, like Kuiper-Belt asteroids, sometimes deeper into the Solar System. They too present a danger to the Earth.

Force and energy: how to change the energy of a body

In this section: we show how the application of a force changes the kinetic energy of a moving body.

The law of conservation of energy strikes us as remarkable when we first meet it because the total energy of an orbiting planet is composed of two parts that each can change, the kinetic and potential energies; only their sum remains constant. This cannot be a magical coincidence: it must reveal something deeper. In this section we see that there is a simple way to see how a force applied to a body (like a planet) changes its kinetic energy.

Let us consider the energy of a planet a little more carefully. Since the potential energy depends only on r , the distance of the planet from the Sun, it follows that the potential energy of a planet on a perfectly circular orbit remains constant in time. From this it follows that the kinetic energy also remains constant. This is the same as saying the speed of the planet is constant, and of course that is what we expect in a simple circular orbit. But the conclusion we are interested in is that the kinetic energy of a planet changes only as its distance from the Sun changes.

Let us see what the change is. Suppose an orbiting planet of mass m moves inwards from a distance r by a very small amount δr to a distance $r - \delta r$. Then in Investigation 6.3 we see that the potential energy changes from $-GmM_{\odot}/r$ to approximately $-GmM_{\odot}/(r - \delta r)$. Since the total energy is constant, the change in the kinetic energy is the opposite (negative) of the change in the potential energy:

$$\text{change in kinetic energy} = \frac{GmM_{\odot}}{r^2} \delta r. \quad (6.18)$$

Notice that the force of gravity (Equation 2.3 on page 13) appears in this expression. In other words we can write this as

$$\text{change in kinetic energy} = F_{\text{grav}} \delta r. \quad (6.19)$$

In other words, the change in energy is the product of the force that moves the planet and the distance the planet moves. In this case, the force is directed inwards, towards the Sun, and we have assumed that the planet also moves inwards, so both the force and change in position are in the same direction. The result is an increase in the kinetic energy. This equation works only for small steps δr in radius. To find the total change in kinetic energy when there is a large change in radius, one must add up successive small changes, in the same spirit as our computer program for the orbit moves the planet in small steps.

Equation 6.19 is quite general, and works no matter what force is applied to a particle. Quite generally, the change in kinetic energy of any particle equals the distance the particle is displaced times the force acting in the same direction as the displacement. If the force acts in the direction *opposite* to the motion of the body, as happens for example with the force of friction as a body slides along a surface, then the change in kinetic energy is negative: we must put a minus sign into Equation 6.19. This corresponds to our expectations: friction reduces the speed and hence the kinetic energy of the body.

Physicists define the right-hand side of Equation 6.19 as the **work** done by the force of gravity. In general the definition is

$$\text{work done by a force} = \text{force} \times \text{distance through which the force acts.} \quad (6.20)$$

As with many common words that physicists use, *work* is not quite the same in physics as in everyday life. Sitting at her desk, a physicist does no work, according

Investigation 6.3. The change in the potential energy

Here we perform a short calculation to find the way the potential energy changes when there is a small change in the distance of a planet from the Sun. The potential energy at the new position $r - \delta r$ can be manipulated with a little algebra into a form that makes it easy to approximate:

$$-\frac{GmM_{\odot}}{r - \delta r} = -\frac{GmM_{\odot}}{r(1 - \delta r/r)} = -\frac{GmM_{\odot}}{r} \left(1 - \frac{\delta r}{r}\right)^{-1}.$$

In this last form of the expression we can use Equation 5.2 on page 43 to approximate the last factor:

$$\left(1 - \frac{\delta r}{r}\right)^{-1} \approx 1 + \frac{\delta r}{r}.$$

When we put this into the expression for the potential energy in the previous equation we find

$$-\frac{GmM_{\odot}}{r - \delta r} \approx -\frac{GmM_{\odot}}{r} - \frac{GmM_{\odot}}{r^2} \delta r.$$

This is the expression we use in Equation 6.18.

Exercise 6.3.1: Changes in potential energy

Justify (or fill in) the algebraic steps that lead from one term to the next in the first equation in this investigation.

to this equation, since she does not change her position. However, no doubt she still expects to get paid for what she does at the desk!

Time and energy

We began this chapter by learning how important the law of conservation of energy is. Then we seemed almost to lose the law, in the gravitational slingshot. Of course, energy conservation does still hold in the slingshot, as long as we add together the energies of both bodies. This is not surprising: we should expect that the spacecraft and the planet could exchange energy with each other. But there is another lesson we can draw from this chapter, and that is about the deep relationship between energy and time.

We looked at two kinds of problems: the motion of a body (planet or spacecraft) around the Sun, and the motion of a body around Jupiter. Both were motions under the action of gravity, and in both cases the body that created the gravitational field was too large to be affected by the body. Yet in one case (the Sun), the total energy of the body was constant, and in the other the total energy changed. The only significant difference between the two problems is that in the first case, the Sun was standing still, and in the second Jupiter was moving. That is, in the first case the gravitational field was time-independent, while in the second the field at any given location depended on time (as Jupiter moved past).

We have here a glimpse into one of the most profound relationships in physics: when there is some underlying time-independence in a physical situation, there is usually a conserved energy, and vice versa. The single body moving past Jupiter does not have a conserved energy because it experiences a force field that is time-dependent. But if we consider the body and Jupiter together, then they move in the background field of the Sun, which is time-independent, so their total energy is conserved.

All the fundamental forces in physics, such as the electric force, gravity, and the nuclear force, work in such a way that the total energy of a collection of bodies is conserved provided that any forces on the bodies from outside the collection are independent of time at any one location.

Essentially, energy is conserved for these bodies if all the rest of the Universe is time-independent. When we come to consider **cosmology** – the study of the Universe as a whole – and the observed expansion of the Universe, we will see how we lose the law of energy conservation: as the Universe expands, its energy simply disappears.

In this section: we make a fundamental and deep connection between energy conservation and time-invariance of the laws of physics.

What about other conservation laws? We have also met conservation of angular momentum and of ordinary momentum. Both of these, as well, are associated with some kind of “independence”. In the case of angular momentum, it is angular independence: the angular momentum of a planet is constant on its trajectory because the gravitational field of the Sun is independent of the planet’s angular position around the Sun. The Sun is spherically symmetric, and this leads to conservation of angular momentum. Ordinary momentum is sometimes called linear momentum because it is conserved in situations where the external forces are constant along straight lines. The ordinary momentum of a planet is certainly not constant along its orbit, and this is because the Sun’s gravitational field is not constant in any fixed direction. But when, for example, billiard balls collide on a billiard table, the effect of the Sun’s (or indeed the Earth’s) gravity is unimportant, and the table itself is flat in any horizontal direction, so the total momentum of the balls in the collision is constant.

Physicists and mathematicians have a name for the general concept that includes time-independence and angular independence. They call it **invariance**. They say that the gravitational field of the Sun is invariant under a change of time (from, say, now to tomorrow), and it is also invariant under a change of angular position. The relation between conservation laws and invariance is something that physicists believe is built into the laws of physics at their deepest level. In fact, physicists today who work on discovering the laws of physics at the highest energies imitate this principle by looking for more abstract kinds of invariances. The approach has been successful so far. Such theories of physics are called *gauge theories*, and all theories of the twentieth century that have unified the nuclear, electromagnetic, and weak forces are gauge theories (see Chapter 27). The principle of invariance is one of the deepest in physics.

Atmospheres: keeping planets covered

There would be no life as we know it on Earth without the atmosphere. Even life in the oceans would not exist: without the atmosphere's thermal "blanket", the oceans would freeze. Yet in the beginning, the Earth probably had a very different atmosphere from its present one. The other planets, with their different masses and different distances from the Sun, all have vastly different atmospheres from the Earth's. In the retention of the atmosphere, and in the subsequent evolution of the atmosphere and of life itself, gravity has played a crucial role.

In this chapter, as we look at the role that gravity has played in this story, we shall encounter fundamental ideas about the nature of matter itself: how temperature and pressure can be explained by the random motions of atoms, why there is an absolute zero to the temperature, and even why atoms cannot quite settle down even at absolute zero. We shall also construct a computer program that builds atmospheres, and we will use it to model not only the Earth's atmosphere, but those of other bodies in the Solar System.

In the beginning ...

The Sun and planets formed some 4.5 billion (4.5×10^9) years ago. We know this from studies of radioactive elements in old rocks, whose decays provide us with a number of natural clocks. The oldest rocks known are older than 4.1 billion years. From theoretical studies of the Sun, which we will describe in Chapter 11, we know that it takes about 4.5 billion years for a star of the Sun's mass to evolve into one that looks like the Sun. It is clear, therefore, that if we want to understand where our atmosphere came from, then we must take a big leap in time-scale, from the orbital periods of planets and space probes we discussed in the last chapter, which are measured in years, to the long perspective of several billions of years.

We do not know a lot about the formation of the Solar System, apart from when it happened. Most likely the planets formed from the same cloud of gas that formed the Sun, material that was not incorporated into the shrinking star, perhaps because it was rotating too fast. How the planets formed from this gas has been made much clearer to us by recent planetary exploration, and particularly by the exploration of the Moon by the astronauts on board the **Apollo** missions.

The first task of explaining the formation of the planets is to account for the great differences between them. The Sun is composed mostly of hydrogen, with some 20–25% helium, and traces of other elements. The giant planets, like Jupiter and Saturn, are also dominated by hydrogen. How, then, is the Earth so solid, with plenty of silicon, iron, oxygen, nitrogen, and other "heavy" elements, but comparatively little hydrogen and helium? Why are all the inner planets rocky and the outer planets gaseous?

All the planets seem to have formed from the hard **dust grains** that pepper the giant **interstellar clouds** of gas. Interstellar clouds are the places where stars form, so their overall composition is like that of the Sun. But, unlike the Sun,

In this chapter: we study the way the atmospheres of the Earth and other planets have developed. We learn how to calculate their structure, and we meet some of the fundamental physical ideas of gases, such as the absolute zero of temperature. We discover the ideal gas law, and we see how pressure and temperature really come from random motions and collisions of atoms. Finally, we look more closely at what happens in a gas at absolute zero, and have our first encounter with quantum theory.

In this section: how the planets formed, and where their atmospheres came from.



they are very cold, cold enough to allow carbon and heavier elements to form tiny condensations that astronomers call dust. These whisker-like grains, only fractions of a millimeter long, are very common in gas clouds, where they are good at blocking the light traveling to us from more distant stars. We shall learn more about where grains come from in Chapter 12.

The gas from which the planets formed contained its share of dust. As a result of random collisions, grains began sticking to each other through molecular forces and building up large lumps. Eventually a number of lumps grew so large that they exerted a significant gravitational pull on their surroundings, pulling in nearby smaller lumps. These are called **planetesimals**. The outer planets (Jupiter, Saturn, Uranus and Neptune) seem to have retained much of the original gas that was in the disk, although they probably started out with rocky cores. Pluto is an exception, and its equally exceptional orbit suggests that it was formed in a different way. Pluto may simply be the nearest large object in the Kuiper Belt.

The **terrestrial planets** (Mercury, Venus, Earth, and Mars) probably trapped gas around themselves initially. But they were close enough to the Sun for this gas to get hot. Since these planets are relatively small, their gravity was not strong enough to hold onto the gas, and the planets lost their initial atmospheres.

In their process of merging from larger and larger bodies, these planets probably experienced their largest collisions last. This helps to account for the fact that many of them spin about an axis that is not perpendicular to the plane of their orbits, and in fact Venus spins in the opposite sense to all the others. The spin is the “memory” of the orbital plane of the last big fragment to merge into the planet. This is also thought to explain our Moon: after the Earth was formed, it was hit by a rogue planet the size of Mars, expelling enough material to form the small “planet” that now orbits the Earth in a repetition in miniature of the formation of the planetary system around the Sun.

The distant planets – Jupiter, Saturn, Uranus, and Neptune – formed in regions where there was more material in the initial cloud and where the temperature was too low to boil off their atmospheres. They may have solid cores but these are hidden from view. They all have moons, some of which may well have been small planets that formed nearby and were captured by three-body collisions. (We will study these in Chapter 13.)

The gas near the terrestrial planets was lost as the planets formed. Only elements trapped in dust grains did not escape. And here is the clue to where the present atmospheres came from: trapped inside the minerals of the grains were not only solid elements, but also some gases, primarily carbon dioxide and nitrogen, with a significant amount of water vapor. These gases were released from the grains when they were heated by the high pressures deep in the interior of the planets. They gradually leaked out of the planets to form the raw material of their present atmospheres. This process is called *outgassing*.

Mercury is hot and small, its gravity too weak to retain an atmosphere at the high temperature to which the Sun heats the planet. So the gases that leaked out simply drifted away, and the planet has no atmosphere. Venus, Earth, and Mars managed to retain small atmospheres. Considering that they all began with similar composition, the fact that these planets have radically different atmospheres now is a striking testimony to the fact that planets *evolve*. The most important factors in their evolution have been geological activity (such as volcanos) and the control of the temperature of the atmosphere by the **greenhouse effect**.

... was the greenhouse ...

Water vapor, carbon dioxide, and methane are **greenhouse gases**: they allow sunlight to pass through to the surface of the planet, but they are opaque to the **infrared** (heat) radiation that the planet radiates back into space. (We will find out more about the greenhouse mechanism in Chapter 10.) They trap this energy near the planet's surface, raising its temperature to well above what it would be on a planet with no atmosphere. Greenhouse gases in the Earth's primitive atmosphere kept it warm enough to have liquid oceans, despite the possibility that the Sun when it first formed was perhaps 25% dimmer than it is today.

On the Earth, we currently worry about the greenhouse effect that might accompany an increase in the small concentration of CO₂ present in today's atmosphere: there is the possibility that human activity will raise the planet's temperature uncomfortably high. Venus, whose atmosphere is more than 95% carbon dioxide, provides an example of an extreme greenhouse effect: Venus is far hotter than the Earth.

The contrast between Venus and the Earth is particularly striking, because they are nearly the same size and at similar distances from the Sun. Venus and the Earth should therefore have started with similar atmospheres. Moreover, both planets have had many volcanos, which release large amounts of carbon dioxide into the atmosphere. Perhaps it is not surprising, then, that Venus has a lot of carbon dioxide. The question is, how has the Earth managed to keep carbon dioxide concentrations low and thereby moderate its greenhouse effect?

In fact, the Earth has just as much carbon dioxide as Venus, but it is no longer in the atmosphere: it is almost all bound up in limestone rocks, which are made of calcium carbonate (CaCO₃). These rocks are made from the deposits of shelled animals, laid down over long periods of time. This shows that, on the Earth, the removal of atmospheric CO₂ has been helped by life itself.

How does this work? When rain falls through the air, some carbon dioxide dissolves in it, and forms in fact *carbonic acid*, H₂CO₃, which is just the combination of a water molecule (H₂O) and one of carbon dioxide (CO₂). When this acid enters the oceans, the bicarbonate ion HCO₃⁻ is freely available to shelled animals, which combine it with calcium that has also been dissolved in the oceans to make more shell, basically more calcium carbonate, CaCO₃. When the shelled animal dies, its shell eventually gets compressed into limestone rocks.

But animals cannot have been involved in this process on the early Earth. Even before shelled animals evolved, the Earth was removing the carbon dioxide released by volcanos, converting carbon dioxide into minerals by chemical means. The bicarbonate ions would form minerals and precipitate out of the oceans if their concentrations were high enough.

What I have described here is a balance that is called the *carbon dioxide cycle*: volcanos put CO₂ into the atmosphere and chemical reactions and living organisms take it out. This cycle plays the crucial role in stabilizing the small amount of carbon dioxide remaining in the Earth's atmosphere, and thus in keeping the greenhouse effect under control. Scientists generally agree on its importance, but the details of how it works are still not clear.

It is remarkable that, at least today, animals help to control the Earth's temperature and keep it fit for life. The attractive possibility that living creatures are actively modifying the Earth's climate in order to make the Earth a suitable place for life is known as the *Gaia hypothesis*: it is hotly debated among scientists today.

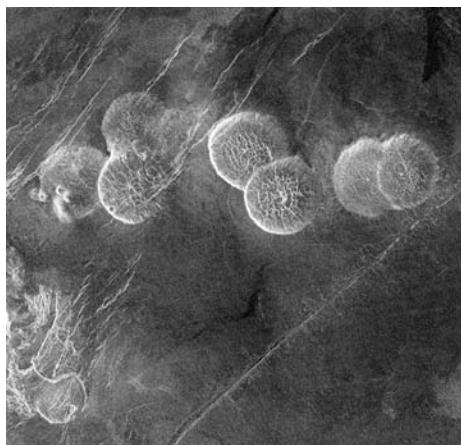
This debate is of more than academic importance. The balance of the carbon

In this section: the sizes of the atmospheres of the rocky planets are very different today. This may reflect the combined effects of differences in the planets' masses and distances from the Sun.

dioxide cycle is now being altered in a possibly dangerous way by human activities that release carbon that had been stored up in coal and oil reservoirs and in forest wood. The total carbon in these stores is small compared to the carbon in rocks, but we are releasing it at a high rate, and (probably not coincidentally) the concentration of carbon dioxide in the atmosphere is rising rapidly. A better understanding of the history of the Earth's atmosphere is now important for making predictions about how the Earth's greenhouse will respond to these human activities.

Equally relevant is the radically different history of Venus' atmosphere. Its original temperature may have allowed liquid water to form oceans, as it did on Earth. But Venus is closer to the Sun than the Earth is, and the carbon dioxide in its original outgassed atmosphere seems to have raised the temperature on Venus high enough to evaporate its oceans. Then the water vapor at the top of the atmosphere was broken up by the Sun's **ultraviolet radiation** into hydrogen and oxygen gas. The hydrogen, being light, escaped. The oxygen, being highly reactive, combined with surface rocks. This gradual depletion of water and the continued release of carbon dioxide by volcanos (see Figure 7.1) resulted in an atmosphere dominated by CO₂, with very little of the original water in it.

Figure 7.1. Venus shows abundant evidence for volcanic activity, including these unusual "pancake" volcano structures imaged by the NASA Magellan mission's radar. Courtesy of NASA/JPL/Caltech.



After Venus lost its water, there was no rain to remove carbon dioxide from its atmosphere, and that sealed its fate. It is intriguing to speculate that perhaps life did actually evolve in the oceans of Venus in its early days, since the planet was not very different from the Earth, and since geological evidence on the Earth suggests that life arose here no later than 1 billion years after the Earth formed, and perhaps even earlier. Life would presumably have been destroyed on Venus when the greenhouse effect evaporated the oceans.

In the end, the difference between Venus and Earth may just be that Venus is a little closer to the Sun, and

its carbon dioxide cycle could not cope with the extra solar energy coming in. If that is the case, what will happen to the Earth if the Sun continues to warm up over the next billion years?

Mars also retains an atmosphere, although it is much thinner than the Earth's. Like Venus, Mars' atmosphere is primarily carbon dioxide. Unlike Venus, where all the CO₂ released by the planet has remained in the atmosphere, most of it on Mars seems to have become locked up in rocks on the Martian surface.

As on the Earth, rain would have removed CO₂ from the atmosphere. Mars shows many geological features that suggest that it had oceans and rivers at one time, for example the channels in the right-hand image in Figure 4.1 on page 26. The conversion of CO₂ to minerals probably occurred through natural chemical reactions, although we cannot exclude the possibility that life evolved on Mars early in its existence and also assisted the removal of carbon dioxide. If this did happen, it was suicide: the big difference between the Earth and Mars is that Mars seems not to have had enough volcanos to replenish the carbon dioxide that was being removed, and thereby to stabilize the greenhouse effect by maintaining a small concentration of CO₂ in the atmosphere.

Volcanos are driven by the release of heat in the interior of a planet. Planets are hot when they form, because of the impacts of all the planetesimals. But planets the size of the Earth and Mars cool quickly, so volcanic activity would go away if there were no further source of heat. For the Earth and Venus, this source is radioactivity.

The decay of naturally-occurring uranium and other elements inside the Earth provides the heat – and the Earth's gravity provides the pressure – that keeps the interior molten. The outward flow of this heat energy drives **plate tectonics**, the drifting and collisions of continents. Plate tectonics keeps volcano activity going. Mars is smaller, so it loses the heat it generates more quickly, leaving its interior temperature too low to drive tectonics and volcanos. Therefore, the size of its atmosphere has steadily declined. This has cooled the planet off: with less CO₂, there is less of a greenhouse effect, and the temperatures decreases.

What happened to the water on Mars? It is probably still all there, frozen into a permafrost. What about life on Mars? If it had time to evolve before Mars cooled off (perhaps 1 billion years) then it froze, or it became starved of CO₂ before that.

...and then came Darwin

As we have seen, the Earth's atmosphere has changed radically since it was first formed. In a combination unique at least to the Solar System, geological, biological, and physical forces have changed the original water–carbon dioxide atmosphere into the oxygen–nitrogen atmosphere we breathe today.

The original life forms evolved in the oceans and were adapted to living with and using the dissolved nitrogen and carbon dioxide from the early atmosphere of the Earth. They used carbon dioxide as their food, and produced free oxygen as a waste product. Gradually, the free oxygen built up to levels that were poisonous to these original organisms, with the result that their closest descendants today, such as anaerobic bacteria, can be found hiding only in rare, oxygen-poor habitats. But new forms of life evolved that actually liked this world-wide pollutant, and from them all the present oxygen-using plants and animals are descended.

The present atmosphere is the result of a balance among a number of forces. I mentioned earlier that the amount of carbon dioxide in the atmosphere has been maintained by a cycle in which chemistry and living things deplete it and natural processes replace it. In all of this, gravity plays a quiet role, regulating the rate of volcanic activity and the rate of atmospheric circulation that leads to weathering. It would be difficult to make an Earth with the same balance of effects if gravity were half its strength or if the planet were twice as massive. Life as we know it can only exist on a geologically unstable planet, and the Earth's size and gravity are perfect for this.

The ones that get away

We saw earlier that hydrogen escaped from Venus' atmosphere because it is light. This has been happening to all the light gases in all the atmospheres of the terrestrial planets: hydrogen and helium have been outgassed, although in much smaller concentrations than in the original cloud of gas, and they have escaped into space.

Does this not contradict Galileo? If gravity makes all things fall at the same rate, regardless of their mass, how can it selectively keep heavier elements in an atmosphere? The explanation is that the *temperature* of the gas is the crucial selecting factor.

Consider a gas with a given temperature, made up of a variety of **atoms** and **molecules**, all with different masses. According to laws first discovered by the Austrian scientist Ludwig Boltzmann (1844–1906), the *kinetic energy* of a typical atom of mass m depends only on the temperature of the gas, and on *nothing else*.

In this section: the evolution of life changed the atmosphere of the Earth, adding oxygen. Life as we know it requires a certain balance of geological activity and atmospheric chemistry. Because of the role gravity plays in these balances, life might not survive on a planet with a very different mass.

In this section: the composition of an atmosphere depends on how many atoms of a given type reach escape velocity. The mean speed of heavier atoms is smaller than lighter atoms, so smaller planets tend to have atmospheres with heavier atoms.

Since the kinetic energy is $\frac{1}{2}mv^2$ (see Equation 6.8 on page 54), the typical speed of an atom in a gas of a given temperature is smaller if the atom's mass is larger.

We shall look more closely at Boltzmann's description of gases later, but it should be clear that not all atoms of a given mass could have exactly the same speed: there is a random distribution of speeds, some much faster, some much slower. It is the *average* speed of each type of atom that will be determined by the gas temperature and the mass of the atom. When a gas is a mixture of several types of atoms and molecules, each type will have the same temperature and hence the same average kinetic energy. Then heavier atoms and molecules in the mixture will have smaller average speeds.

Now we can see how gravity affects an atmosphere. The deciding factor for whether or not any body is bound to another by gravity is whether its speed exceeds the *escape speed* for that particular body. Since light atoms will have a higher speed than heavier atoms, they will have a greater tendency to escape. Those atoms which have by chance a random speed much higher than the average speed for the prevailing temperature will escape sooner. Although some atoms of any mass will always escape, there will be a critical mass above which so few atoms have enough speed to escape that atoms and molecules of that type and those that are more massive will be effectively bound to the planet.

Since more massive planets have larger escape speeds, they will retain more of their lighter atoms and molecules than less massive bodies can. Massive planets far from the Sun at low temperatures have retained all their light gases, like hydrogen. Smaller planets near the Sun have atmospheres that consist primarily of heavier atoms and molecules.

Atoms that are light enough to escape from a planet can be present in its atmosphere only if they are constantly replenished. Helium, for example, exists in the Earth's atmosphere mainly because it is generated by radioactivity in rocks: some radioactive elements produce alpha-particles, which are nuclei of helium. When these particles slow down and come to rest in rocks, they acquire two electrons and become helium atoms. They are not chemically bound to the rocks, so eventually they escape into the atmosphere, where they stay for a geologically short time before escaping into space.

The Earth's atmosphere

In this section: to understand even the simplest aspects of the structure of the Earth's atmosphere, we must learn about how pressure and gravity balance.

The Earth's atmosphere is an extremely complex system. It has many distinct layers; it is subject to continual mixing caused by weather systems and by **convection** of hot air from the ground; it is dragged along with the rotation of the Earth; it receives large amounts of water vapor over the oceans and dumps much of the water on the continents; it is heated by the Sun at a number of different altitudes, depending on which radiation-absorbing gases are where; and its composition is constantly evolving from the actions of natural forces and of man.

A full discussion of these complexities is well beyond our scope here, and in fact is well beyond the scope of any present computer model of the atmosphere. The vast number of effects that need to be taken into account would overwhelm the speed and memory of even the biggest supercomputers today, and in many cases the complexity of the physics and chemistry of the atmosphere defies understanding at present. Nevertheless, computer models of the atmosphere are helpful and informative if they are interpreted with due care.

For us, the computer can again be an aid to solving even the simplest equations that describe atmospheres. Our aim here is to understand enough of what affects an atmosphere to be able to use a computer to model the Earth's static atmosphere, where we neglect any changes due to weather, circulation, and so on.

The fundamental point is that the structure of the atmosphere is essentially a balance between gas pressure, which pushes the atmosphere up, and gravity, which keeps it down. We already know enough about gravity to write our program. But we do not know enough about pressure yet. We need to discuss two points about pressure that are crucial to the construction of a simple computer program. The first is how pressure manages to push things up; the second is the way that temperature affects pressure. Only after we discuss these will we be in a position to see how gravity and gas pressure balance each other to determine the structure of the Earth's atmosphere.

Pressure beats gravity: Archimedes buoys up balloons

A good way to understand how pressure acts to keep an atmosphere up is to see how it pushes a helium-filled balloon up through the atmosphere. We all know that the balloon rises because helium is lighter than air: at a given pressure and temperature, helium atoms weigh considerably less than the nitrogen and oxygen molecules of the surrounding air. But how, in detail, does the balloon rise? Where are these **buoyancy** forces acting, and where do they come from?

It is clear that the forces must come from the pressure forces in the surrounding air: the balloon is not in contact with anything else that can exert forces on it. Scientists define the pressure on any surface, such as a spherical balloon, to be the *force per unit area, acting perpendicular to the surface*. So on each little patch of balloon surface, there is a pressure force pointing inwards toward the center of the balloon. The *pressure force* on the patch equals the *pressure* at that place on the balloon times the area of the little patch.

Now I will pose an apparent paradox. Pressure is an **isotropic** force. This means that at any point inside a gas, pressure pushes with the same force in all directions. In other words, it doesn't matter whether I position the balloon just above the point or just below it, just to the left or to the right of the point in question, the pressure force will be exactly the same.

How, then, can the balloon rise? After all, a balloon has weight, so gravity is pulling it down. For it to rise, there must be another force pushing it up even more strongly. How can pressure provide this force? Doesn't the pressure of the gas above it push it down with a force just equal to the pressure of the gas below it pushing up? If this is so, don't the two forces exactly cancel?

The answer is, of course, no. The reason is that the pressure above the balloon is not acting *at the same point* as the pressure below the balloon: it is acting on the other side of the balloon. Therefore the isotropy of the pressure is no reason to expect the two forces to cancel each other exactly, provided the pressure changes from one *place* to another. We say that pressure is isotropic, but it is not homogeneous: it is not the same at all different locations in the atmosphere.

In fact, pressure goes down as we go higher in altitude. This is the reason for the familiar sensation in our ears if we change altitude too quickly: the air pressure inside our heads does not change as quickly as that outside, and the inequality of pressure on the two sides of the eardrum causes a painful strain on that delicate membrane. So the balloon has less pressure on its top than its bottom, and the net effect of pressure is to push upwards on the balloon.

It is only the *non-uniformity* of pressure across an object that provides a net pressure force on it.

But of course this is only part of the story. The same thing happens to, say, a rock held in the air: the pressure above it is lower than the pressure below it. Yet it

In this section: we learn exactly what the combination of forces is that lifts balloons into the air, and what needs to be balanced for the atmosphere to be in a steady state.

will fall when let go, not rise. For the balloon to rise, we have to have that the net upwards pressure force is actually larger than its weight.

Where is the dividing line? What is the critical weight that an object must have so that gravity and the pressure difference across it just balance and allow it to remain at rest? To answer this, let us try to find something that does remain at rest when it is let go, something that is said to be *neutrally buoyant*. The simplest example of such a thing is air itself. The air that is displaced by the balloon when we place it in the atmosphere, the air that would otherwise occupy the same place as we have put the balloon: this air would not move. So it has just the right weight to remain at rest. Therefore, if the contents of the balloon are *lighter than air*, like helium, the balloon will rise. If the object is heavier, like a rock, it will fall.

The motion of an object in an atmosphere is the result of the net force that results when the pressure force down on it from above, the pressure force up on it from below, and the downward force of gravity on it are all added together.

Now we can return to the question we started with in this section: how does pressure support the atmosphere? Our example of neutral buoyancy shows how: if we replace our helium-filled balloon with the parcel of air again, then we see that it is held up against gravity by the pressure difference across it:

For the atmosphere to be in equilibrium (in other words, perfectly at rest) the pressure force at the bottom of any parcel of air must exceed that at the top by the weight of the parcel.

►The word “hydrostatic” reveals how scientists began thinking about buoyancy: by studying things floating on water. The prefix *hydro-* is from the Greek word for water.

►The fact that pressure is isotropic, exerting forces in all directions, is what makes hydraulic machines work: the tubes that hold the brake fluid in an automobile braking system transmit the pressure from the brake pedal to the wheels regardless of twists and turns in the line. Pascal invented the hydraulic press, so we owe our automobile braking systems to his insight. He also founded the mathematical theory of probability, so we owe the calculations of our automobile insurance premiums to his insight, too!

In this section: we explain how even objects that are heavier than air can use air pressure to keep them aloft.

The mathematical expression for this rule is Equation 7.1 of Investigation 7.1, which is called *the equation of hydrostatic equilibrium*. The neutral-buoyancy argument of this section, which tells us that objects immersed in a fluid (such as air or water) will sink if they weigh more than the fluid they displace, was first worked out by the Greek scientist and engineer Archimedes (about 287–212 BC).

A very simple application of the law of hydrostatic equilibrium is to the whole atmosphere. Consider a slender column of air, stretching from the ground to the top of the atmosphere. The pressure at the top is zero, so the pressure force difference from top to bottom is the pressure force at ground level, and this is the pressure there times the cross-sectional area of the column. This must exactly balance its weight.

The atmospheric pressure on any area of the ground must be large enough to support the entire weight of the column of air above that area. Since atmospheric pressure is measured to be $1.013 \times 10^5 \text{ N m}^{-2}$, the weight of the air above a balloon of radius 10 cm and cross-section 0.0314 m^2 is 3140 N, which by Equation 2.2 on page 10 is the weight of a mass of about 320 kg, or four heavy men.

The units for pressure, N m^{-2} , are given a standard name, the pascal, denoted Pa. This is named after the French physicist and mathematician Blaise Pascal (1623–1662), who first pointed out the isotropy of pressure.

Pressure beats gravity again: Bernoulli lifts airplanes

Given that atmospheric pressure can support four heavy men, it is perhaps not so surprising that the atmospheric pressure difference across a small helium balloon can be enough to give the balloon a small push upwards. But balloons are small-fry: atmospheric pressure really shows its strength when it lifts airplanes.

Investigation 7.1. The balance of pressure and gravity in an atmosphere

Here we shall translate into equations the words about hydrostatic equilibrium at the end of the section on buoyancy. We fix our attention on a certain cubical parcel of air in the middle of the atmosphere, whose imaginary boundaries are drawn simply to distinguish it from the rest of the atmosphere. To remain at rest, the net force on the gas in it must vanish, regardless of the fact that its sides are imaginary. Newton's laws apply to any collection of particles, even if they do not form a solid body.

We take the length h of the sides of the cube to be small enough that there is only a small change of pressure across the cube, and (consequently) that the density of the air inside the cube is essentially uniform.

The three forces that affect the vertical motion of the cube's contents are the pressure at the bottom, the pressure at the top, and the force of gravity (the weight of the parcel of air). Let us consider them in turn.

Pressure force on the bottom of the parcel. If we denote the pressure at the bottom by p_{bottom} , then the force on the bottom is the pressure times the area of the bottom, which is a square of side h :

$$F_{\text{bottom}} = p_{\text{bottom}} h^2.$$

Pressure force on the top. Let the change in the pressure going from the bottom to the top be Δp ; since the pressure falls with height, we expect this will come out to be a negative number. Then the pressure at the top is $p_{\text{top}} = p_{\text{bottom}} + \Delta p$, and the pressure force on the top of the cube is

$$\begin{aligned} F_{\text{top}} &= -(p_{\text{bottom}} + \Delta p)h^2 \\ &= -p_{\text{bottom}} h^2 - \Delta p h^2. \end{aligned}$$

Here the overall minus sign is needed because this force points *downwards*, in the direction opposite to the force on the bottom.

Force of gravity. The weight of the parcel is its mass times the acceleration of gravity, g . The mass, in turn, is the volume h^3 of the parcel times the density of air inside the parcel, which we call ρ .

Exercise 7.1.1: How fast does a helium balloon rise?

The density of the helium in a balloon filled at, say, a fairground is 0.18 kg m^{-3} , while the density of the air around it is 1.3 kg m^{-3} . Using Equation 7.1, compute the pressure difference across the air that the balloon will displace, assuming for simplicity that it is a *cube* of side 20 cm. Then compute from this the net pressure force on the balloon itself when it is inflated and takes the place of the air. Next, compute the weight of the balloon (neglecting the rubber of the balloon itself) and calculate its initial acceleration (upwards). What multiple of g is this? Will it keep this acceleration as it moves upwards? How many balloons would be required to lift a 60 kg woman?

Since this force also points downwards, we have

$$F_{\text{gravity}} = -g\rho h^3.$$

Net force. These three forces must add up to zero for the atmosphere to be in equilibrium. The equation is

$$\begin{aligned} 0 &= F_{\text{bottom}} + F_{\text{top}} + F_{\text{gravity}} \\ &= p_{\text{bottom}} h^2 - p_{\text{bottom}} h^2 - \Delta p h^2 - g\rho h^3 \\ &= -\Delta p h^2 - g\rho h^3 \end{aligned}$$

This can be solved for the pressure step going from bottom to top:

$$\Delta p = -g\rho h. \quad (7.1)$$

This is called the *equation of hydrostatic equilibrium*, written in terms of *differences* between the top and bottom of the parcel. It is accurate only as long as h is sufficiently small, since the density ρ used in the equation is the average density over the parcel, and there would be errors if the density changed much across the parcel.

Readers who know calculus will recognize the beginnings of a standard limit argument. (Others should ignore this paragraph and the next!) As h tends to zero, so does Δp . Since h is really the change of altitude z in the atmosphere, it is convenient to change notation and replace h by Δz . Then dividing Equation 7.1 by Δz and taking the limit gives

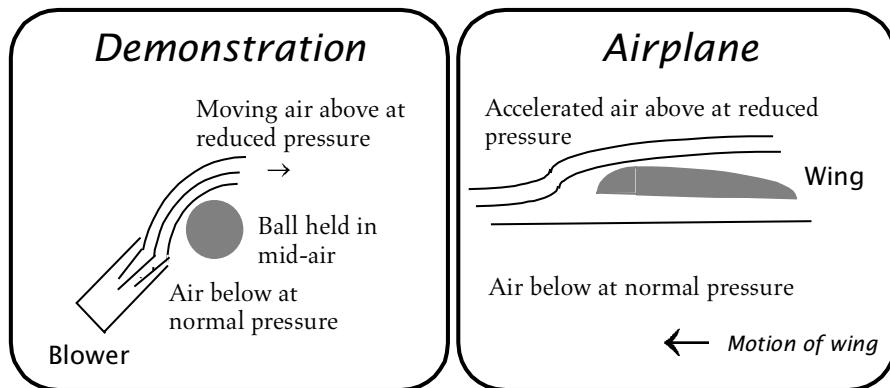
$$\lim_{\Delta z \rightarrow 0} \left(\frac{\Delta p}{\Delta z} \right) = \frac{dp}{dz} = -g\rho.$$

This is the calculus version of the equation of hydrostatic equilibrium, the version that most physicists would think of if asked to write down the equation. But our finite-difference version, Equation 7.1 above, is the starting point for physicists when they actually solve the equation on a computer. We shall build our computer program for constructing model atmospheres on the finite-difference version, which contains the same physics as the differential equation.

Simple buoyancy is not enough to lift a airplane. Being much heavier than the air it displaces, an airplane will not float upwards the way a dirigible does. To fly, an airplane creates an artificial kind of buoyancy: it uses its *speed* to create a greater pressure difference between the upper and lower surfaces of its wings than there is in the resting atmosphere. We have learned enough about Newton's laws now to be able to understand how this works.

The basic effect was first observed by the Swiss physicist Daniel Bernoulli (1700–1782): if a fluid speeds up to get past an obstacle, its pressure goes *down*. The reason is simply Newton's second law. To speed the gas up, there must be a net force on it in its direction of motion. A small piece of the fluid that moves around the obstacle does not touch the obstacle, so the only forces it experiences are pressure forces from the surrounding fluid. The acceleration must, therefore, ultimately come from the net pressure forces in the fluid. For pressure to give a net force in the direction of motion, the pressure must be larger behind the bit of the fluid we are considering, and smaller ahead of it. Therefore, when the fluid accelerates to pass around an obstruction, its moves from a region of higher pressure to one of lower: its pressure drops. When it passes the obstacle, it rejoins the flow of the rest of the fluid, so its

Figure 7.2. Illustrating the Bernoulli effect as it is often demonstrated in science exhibitions (left) and as it works to lift an airplane (right).



speed drops. To slow it down, the pressure must rise as it moves further forward. The minimum pressure will therefore be found at the place where the speed of the fluid is greatest.

Many readers will have seen a popular demonstration of the Bernoulli effect in a science museum or exhibit. It consists of a jet of air directed at a large inflated ball, which causes the ball just to hang in mid-air (see Figure 7.2). The ball hangs where the air from the jet will strike it just above its center and pass over it. This reduces the pressure of the air above the ball, increasing its buoyancy and stopping it from falling. This artificial buoyancy is called *lift*.

The wing of an airplane is shaped to make use of the same effect. It is rounded on top and flat below. As the airplane's engines push the airplane through the air, the air that passes above the wing has further to go, since the curved surface is longer than the flat one. The air passing above the wing accelerates, and its pressure drops. This provides the lift that the airplane needs to fly.

Pilots change the lift on a wing by modifying its shape with wing flaps. In this way they can provide enough lift at low speeds, while taking off or landing. Nature knew about the Bernoulli effect, of course, long before scientists did. Just like pilots, gliding birds shape their wings to take advantage of the same physics.

Helium balloons and the equivalence principle

Our discussion of buoyancy has a lovely link with our earlier study of the equivalence principle. The buoyancy of a helium-filled balloon gives rise to a classic "trick" physics problem, which we have all the preparation necessary to solve correctly. Try this experiment with your friends, but be sure the road is empty when you do!

Suppose you are in a car moving at a constant speed with the windows shut and the vents closed, and you are holding a helium-filled balloon by its string. The balloon keeps the string vertical. Now the car brakes hard. Which way does the balloon move: towards the front or the back of the car?

Are you ready for the answer? Consider the equivalence principle, which in its simplest form says that a uniform acceleration produces the same effects as a gravitational field in the opposite direction. So the environment inside the car when it brakes would be the same if we took the uniformly moving car and added a gravitational field pointing towards the front. Then the buoyancy of the balloon causes it to rise against gravity: in this case, to move towards the *back* of the car.

►By extending wing flaps before landing, the pilot insures that air passing over the wing has further to travel, and therefore that it goes faster and provides more lift.

In this section: how to use the equivalence principle to solve a difficult physics problem involving balloons.

If you don't like to use the equivalence principle, you can work it out in terms of the pressure forces we discussed earlier, since the pressure of the air in the car will be higher at the front than at the back as the car decelerates. But you will still have to find something to play the role of the "weight" of the balloon when you work out the net horizontal force, since a rock would certainly not move towards the back of the car, despite the pressure difference. You are welcome to elaborate this argument (and of course a good physicist should be able to do so), but for my taste the equivalence principle provides a much more elegant approach to the solution.

Absolute zero: the coldest temperature of all

Buoyancy is only half of the story of how an atmosphere is supported. Now we turn to the other half. The temperature of an atmosphere is important because hotter gases tend to expand and have a lower density at a given pressure, so they are more buoyant.

What is the relation between temperature and volume? If a gas is cooled under fixed pressure, experiment shows that its volume decreases in direct proportion to its temperature; if we make twice as large a change in its temperature, its volume decreases by twice as much. Expressed as a word equation, this is

$$\text{change in volume} = \alpha \times \text{change in temperature}, \quad (7.2)$$

where α is a constant that will depend on the pressure. This is called *Charles' Law*, named for the French physicist and mathematician Jacques-Alexandre-César Charles (1746–1823), and it has a remarkable consequence.

Consider a finite volume V of the gas. If we keep reducing the temperature, eventually the decrease in volume in Equation 7.2 will equal V : the volume will have shrunk to zero! By Equation 7.2, this has required only a *finite* lowering of the temperature. The temperature at this point is surely the coldest one can make that gas. Therefore Charles' law implies that every gas has a *coldest* temperature.

Even more remarkable is that we can easily show that this coldest temperature is the *same* for every gas, no matter what pressure or volume we start with. Recall an everyday experience with temperature: if we place ice cubes into a warm drink, we get a cooler drink in which the ice has melted. In general, if we put two bodies that have different temperatures into contact, they gradually approach a common temperature somewhere in between the two. The warmer one cools and the colder one warms.

Now suppose the two bodies are two gases that have both been cooled to their own coldest-possible temperatures. If these temperatures are different, and we bring them into contact, then the warmer one will cool further. But this is a contradiction: it is already as cold as it can be. Therefore, our assumption that the coldest temperatures were different was wrong: the coldest temperature must be universal.

Charles' Law implies that there is a universal lowest temperature. Experiment reveals that this is about -273°C or -460°F . It is natural to define a new *absolute* temperature scale for which this temperature is defined to be zero. We call the coldest temperature *absolute zero*.

The absolute temperature scale (whose degree size is the same as the **celsius** scale) is called the **kelvin** scale, and temperatures are denoted by "K". The freezing point of water is about 273 K, and the boiling point of water is about 373 K. The advantage of using this scale is that, at constant pressure, the volume of a gas is zero when the temperature is zero, so Charles' law can be re-phrased to say that the volume is directly proportional to the absolute temperature. Thus, Charles law tells us that, if a given gas has volume V at atmospheric pressure and at absolute

In this section: by considering how the volume of a gas changes with its temperature, we are led to conclude that there must be an absolute zero of temperature.

►Charles also understood buoyancy and was prepared to stake his life on it: he was the first person to ascend in a hydrogen balloon, reaching heights well over 1 km.

►It is conventional in scientific writing today just to write "K", not the old-fashioned " $^{\circ}\text{K}$ ", after a temperature measurement.

In this section: we learn about one of the fundamental achievements of physics, namely the understanding that temperature measures the energy of the random motions of atoms. This insight is fundamental to understanding atmospheres and, in later chapters, stars.



Figure 7.3. Ludwig Boltzmann was one of the giants of physics at the transition from the nineteenth to the twentieth century, when the foundations of modern physics were laid. In Boltzmann's time, not all physicists accepted that matter was composed of atoms. Many of the ideas Boltzmann used go back to Bernoulli, who argued that pressure could be caused by the random motion of small particles, and that this would naturally explain the increase of pressure with temperature. But Boltzmann put mathematics to these ideas, and turned them into a testable physical theory. By doing this, Boltzmann made a great step toward establishing the reality of atoms. Unfortunately, Boltzmann's story is a good deal more tragic than that of most other physicists: in his middle age, ill and depressed, at least partly by the resistance his ideas had met among older physicists, he committed suicide. He never understood that he had completely converted the younger generation of physicists to his point of view, and he did not know that his theories were actually on the verge of experimental verification.

Image courtesy Österreichische Zentralbibliothek für Physik.

temperature T , then at absolute temperature $2T$ and the same pressure it will have volume $2V$.

Why there is a coldest temperature: the random nature of heat

Although the universal nature of the absolute zero of temperature has been verified over and over again in laboratory experiments, there might be something unsatisfying about our approach to it so far: we have no real explanation, no real understanding of how this can be. A more satisfying explanation was provided by Boltzmann, whom we mentioned earlier in this chapter.

Boltzmann was the principal founder and exponent of the branch of physics that we now call **statistical mechanics**. (Other important contributions were made, independently, by James Clerk Maxwell – whom we will meet in Chapter 15 – and by the American physicist Willard Gibbs, 1839–1903.) Boltzmann showed that *all* the known properties of simple gases could be explained if one took the view that a gas was composed of atoms that moved randomly about inside a container, frequently colliding with each other and with the walls of the container. He showed that pressure was the result of the forces of all the small atoms hitting the walls randomly. To make his calculation work, he needed to make only one simple assumption about the relationship between the average kinetic energy of an atom in the gas[†] $\langle K \rangle_{\text{avg}}$ and the absolute temperature:

$$\langle K \rangle_{\text{avg}} = \left\langle \frac{1}{2}mv^2 \right\rangle_{\text{avg}} = \frac{3}{2}kT, \quad (7.3)$$

where we have used Equation 6.8 on page 54, the definition of the kinetic energy for an atom of mass m . The constant k is called *Boltzmann's constant*, and it has the value

$$k = 1.38 \times 10^{-23} \text{ kg m}^2 \text{ s}^{-2} \text{ K}^{-1}.$$

Equation 7.3 is the quantitative form of the relation between temperature and kinetic energy that I referred to at the beginning of this chapter. We study Boltzmann's argument in more detail in Investigation 7.2 on page 78.

The idea that kinetic energy should be proportional to temperature was not just an arbitrary assumption. What Boltzmann showed was that when a large collection of atoms move and collide randomly, they tend to share out their kinetic energy equally: when a rapidly moving and a slowly moving atom collide, they usually both bounce off with speeds somewhere in between. This is so similar to what happens when hot and cold bodies are placed into contact, that Boltzmann drew what was to him an obvious conclusion: temperature essentially *is* the kinetic energy of a typical atom of the gas.

This leads, of course, to a simple explanation of *why* bodies in contact tend to approach the same temperature: their atoms at the point of contact tend to share energy, and when they collide with atoms behind them inside their respective bodies, this sharing tends to make all energies – hence both temperatures – the same. Moreover, and this is where our real interest is in this section, Boltzmann gives us a natural explanation for absolute zero: absolute zero is the temperature at which there is no longer any random kinetic energy inside the body. At absolute zero, all the atoms are perfectly at rest with respect to each other. The fact that this lowest temperature should be the same for all bodies is obvious in this picture.

Why does absolute zero lead to zero volume? Remember that in Charles' law the pressure is held constant, so there is always some pressure from outside on the

[†]The use of angle brackets $\langle \dots \rangle_{\text{avg}}$ is a conventional notation for a statistical average (also called the **mean**) over a large number of random events. In this case the average is over random motions of molecules.

gas. As its temperature decreases, the random motions of its atoms get slower, and their ability to resist compression decreases, so the volume decreases. Ultimately, when the atoms stop moving, they have no resistance to compression at all, and the volume goes to zero.

Notice that temperature is related to the *random* kinetic energy of the atoms. If we take a body at absolute zero and make it move at a constant speed, each atom will have a kinetic energy, but there will be no random motion: all the atoms are at rest with respect to one another. So the temperature will still be zero.

Although we have characterized absolute zero as a state in which the atoms of the gas stop moving, this state cannot actually be reached: no matter how one tries to remove kinetic energy from a gas, one will always do something to disturb it a little and leave a small amount behind. This may be very small, so one may try to get as close to zero as one likes; but absolute zero is unattainable. To date, temperatures below 0.001 K have been reached in small samples, and a metal bar weighing more than a ton has been cooled to below 0.1 K. (This bar has another connection with gravity: it is used in a gravitational wave detector, which we will discuss in Chapter 22.)

These low temperatures may be the lowest ever seen anywhere in the Universe. We will see later that the Universe began as a hot gas (the Big Bang), and has been cooling off ever since. But there is a background of stray radiation left from the Big Bang that keeps the temperature of all natural objects above a minimum of about 2.7 K. To get colder than that probably requires some deliberate intervention. If the Earth contains the only intelligent life in the Universe, then cold temperatures may have existed only here.

The ideal gas

We have studied the way the volume of a gas depends on its temperature, but in doing so we held the pressure constant. We must now ask about changes in pressure.

In Boltzmann's picture, it is clear that, if we fix the volume of a gas and reduce its temperature to absolute zero, then the random motions of atoms go to zero, and the pressure (which results from the impacts of gas atoms on the walls of the container) must also go to zero. Conversely, as we raise the temperature at constant volume, we should expect the pressure to rise. Boltzmann showed by calculations what experiment had already confirmed: that the pressure is directly proportional to the temperature in these circumstances.

These two laws can be combined into a single relation, which is called *the ideal gas equation of state*: the absolute temperature of a gas is proportional to the product of pressure and volume. This is expressed mathematically as:

$$pV \propto T. \quad (7.8)$$

This is the key relation for seeing how the density and pressure of the atmosphere change as we go up in altitude (see Figure 7.4 on page 79). In Investigation 7.2 on the next page we derive this relation from Boltzmann's point of view. We find there that the constant of proportionality in this equation is just Nk , where N is the total number of atoms in the gas and k is Boltzmann's constant.

Our work in Investigation 7.2 also gives us another important relation, namely that the typical velocity v of atoms in the gas can be found just from its density ρ and pressure p . Since sound waves in a gas are nothing more than some atoms

In this section: the simplest gas consists of independent atoms that collide with one another but do not stick together or lose energy through collisions. We learn that the pressure, volume, and temperature of such a gas have a simple relationship to one another, and that the sound speed depends on the ratio of pressure to density.

▷ The symbol “ \propto ” stands for “is proportional to”.

Investigation 7.2. The ideal gas according to Boltzmann

Before Boltzmann, scientists understood two simple relations among the pressure p , volume V , and absolute temperature T of a gas: $V \propto T$ with p fixed (Charles' law), and $p \propto T$ with V fixed. By multiplying these two equations, one gets a single relation among all three quantities:

$$pV = \beta T, \quad (7.4)$$

where β is a constant, independent of p , V , or T . Boltzmann showed how to find β in terms of the atoms that make up the gas. We sketch his argument here.

Boltzmann observed that pressure represented the force that the gas exerts on the walls of its container, and by Newton's third law (Chapter 2) this is the same as the force exerted by the walls back on the gas. How does this force act? If the gas consists of atoms, then any wall will exert a force only while atoms are in contact with it, bouncing off it. The result of the force on the atom is to turn it around, to reverse the component of its velocity that is perpendicular to the wall. If we had to know the details of this process, such as how long it took the atom to turn around, how hard or soft the wall was, and so on, then the calculation would be hopelessly complicated.

Luckily, we are only interested in the *average* force exerted by the wall, so we can make a simplification: if the typical time between the collision of one atom with the wall and the collision of the next atom is Δt , then the average force exerted by the wall is the same as would be required to turn a single atom around during the time Δt . The reason is that, if the time it takes an atom to turn around is longer than Δt , then the force exerted on each atom will be less than we calculate, but at any time there will be several atoms being turned around, so the total force exerted by the wall will be the same as if one atom were turned around in the time Δt . A similar argument shows that this is the right *average* force when the time to turn an atom around is shorter than Δt , as well.

We shall now calculate how the pressure force depends on the atoms' average speed. If the average speed of an atom is v , then the component of a random atom's velocity perpendicular to the wall is proportional to v . The acceleration experienced by an atom turned around in time Δt will be the change in its velocity divided by Δt , so this will be *proportional to $v/\Delta t$* . If the atoms have mass m , then the pressure exerted by the wall will satisfy

$$p \propto mv/\Delta t.$$

Now we need Δt . The time between successive collisions of atoms with a wall will certainly depend on v : the faster the atoms travel, the smaller will be the interval between collisions. It also depends on the number of atoms per unit volume: the more atoms there are within a given distance of the wall, the more collisions there will be. The number per unit volume is the total number in the container, N , divided by its volume, V . We have thus argued that Δt is proportional to $1/v$ and to $1/(N/V)$, or that $\Delta t \propto V/Nv$.

When we put this together with the previous equation, we find

$$p \propto mv^2 N/V. \quad (7.5)$$

Exercise 7.2.1: How many atoms in a balloon?

Consider the cubical helium-filled balloon of Exercise 7.1.1 on page 73. If the pressure inside the balloon is atmospheric pressure, $p = 10^5 \text{ N m}^{-2}$, and the temperature is $T = 300 \text{ K}$ (about 81 F), then use Equation 7.6 to calculate the number N of helium atoms in the balloon. The size of this answer justifies the approximation that we can average over large numbers of randomly moving atoms.

Exercise 7.2.2: What is the mass of a helium atom?

Use the answer to the previous exercise and the density of helium given in Exercise 7.1.1 on page 73 to calculate the mass of each helium atom. Use the density given for air to calculate the *average* mass of an air molecule. (Since air is a mixture of gases, we only obtain the *average* mass this way.)

Multiplying by V and using Equation 7.3 on page 76 to replace the typical value of v^2 by something proportional to the temperature T , we find

$$pV \propto NkT.$$

This is equivalent to Equation 7.4, and it tells us one more thing, namely that the constant β in that equation contains Nk , the total number of atoms in the gas and Boltzmann's constant.

Now, Boltzmann was able to do the calculation better than we did, because he was careful to do the averages over all the directions of the actual velocities of the atoms in the gas, so he could calculate the constants of proportionality in each of the steps. For his definition of k , as given in Equation 7.3 on page 76, he showed that the constant of proportionality in the above equation was just one:

$$pV = NkT. \quad (7.6)$$

The argument we have given shows that this equation holds for all gases: atomic hydrogen, molecular oxygen, and inert helium all obey Equation 7.6. This is called the *ideal gas equation of state*.

This equation is "ideal" because we have made an oversimplification by assuming that the atoms interact with one another and with the walls only when they actually collide. In real gases, there can be various electric forces between atoms even when they are well separated, that make slight modifications in this equation. The real exceptions come when the gas becomes a liquid (or worse, a solid), as most gases do at low enough temperatures. Then the interactions between atoms become very strong, and the system cannot even be approximated as one composed of free atoms colliding occasionally.

One consequence of our derivation is an expression for how the pressure depends on the average kinetic energy of the atoms. By replacing kT in Equation 7.6 by $2\langle KE \rangle_{\text{avg}}/3$, we find

$$p = \frac{2}{3} \frac{N}{V} \langle KE \rangle_{\text{avg}}. \quad (7.7)$$

We will use this equation when we study stars.

Another look at Equation 7.5 will show how, by making measurements on a gas, we can deduce the speed of its atoms. The quantity mN/V in the right-hand side of this equation is just the density ρ of the gas, the total mass per unit volume. So we learn that $v^2 \propto p/\rho$.

The ideal gas equation of state gives another perspective on "why" the helium-filled balloon rises. Both the balloon and the air it replaced had to have the same pressure, since they were surrounded by air with that pressure. They had the same volume and temperature, too, so they must therefore have had the same number N of atoms (or molecules, in the case of air). The force of gravity on the helium balloon is less because each atom of helium is so much lighter than an average molecule of air (which is a mixture of molecules of nitrogen, N_2 , oxygen, O_2 , and other gases).

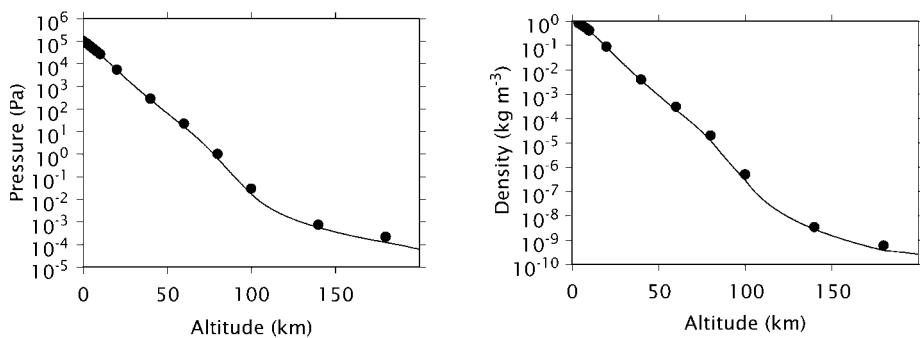


Figure 7.4. Comparison of predictions of the computer model of the Earth's atmosphere (solid lines) with the measured data (points).

bumping into others and making them bump into ones further away, the speed of sound is also given just by the pressure and density:

$$v_{\text{sound}}^2 \approx p/\rho. \quad (7.9)$$

In our discussion of Boltzmann's picture of a gas as composed of atoms bouncing around, we did not really talk about what the atoms are. For Boltzmann, they were just little particles that somehow characterized the gas. In his gas laws, the "atoms" are whatever fundamental units the gas is composed of. Thus, if the gas consists of single atoms, as in helium gas, then the particles are the helium atoms. But if the gas is a molecular gas, such as oxygen, which normally exists as O₂, then Boltzmann's laws apply to the molecules. For example, the number N in the constant of proportionality in Equation 7.8 on page 77 would be the number of O₂ molecules, not the number of oxygen atoms.

An atmosphere at constant temperature

Imagine a column of air above a square drawn on the Earth. Let us go up from the Earth a small height, perhaps a few centimeters, so that there are N air molecules above the square to that height. Now imagine marking off successively higher steps, each of which makes a volume that contains the same number N of molecules. (These steps are not generally equally spaced in altitude, of course, because the density is decreasing – the air is getting "thinner".) If the air is still or moving slowly, then the forces on it must be in balance. What are these forces?

First, what is the gravitational force? We shall assume that the atmosphere does not extend very far from Earth, so that the acceleration of gravity g is constant everywhere inside. This is not a bad assumption, since the top of the atmosphere is certainly within 300 km of the ground, which is the altitude where many satellites orbit. This is less than 5% of the radius of the Earth, so to a reasonable approximation we can neglect the weakening of gravity as we go up. Then the gravitational force will be the mass of each volume times g .

Next we need to calculate the pressure forces. In order to be in equilibrium, the pressure force on the bottom of each volume must exceed the pressure force on its top by the weight of the molecules in the volume. Since each volume contains the same number of molecules, this weight is the same for each volume; and since we have constructed our volumes to have equal areas, the pressure change from one step to the next must be the same, all the way up. When the pressure falls to zero, we are at the top of the atmosphere.

To calculate the pressure changes, we have to have information about the way the temperature changes with height. In this section, we make the simplest assumption: we consider only the constant-temperature, or **isothermal**, atmosphere.

In this section: the simplest atmosphere to study is one with a uniform temperature. We show that it is of infinite extent: it cannot have a top boundary. Therefore, real atmospheres must have non-uniform temperatures that fall to zero at the top.

▷We assume, of course, that the temperature is not absolute zero!

Although not an entirely realistic representation of what happens on the Earth, it is nevertheless an instructive first example to think about, because it is easy to see what happens. Besides, as we shall see, this situation does arise in portions of other planetary atmospheres.

For a constant temperature, the volume of a gas is inversely proportional to its pressure. As the pressure goes down from one step to the next, the volume must go up in proportion. Now, since the cross-sectional area of each volume is the same, this means the height of each step increases in inverse proportion to the decreasing pressure. How far up do we have to go to get to the top of this atmosphere?

The answer is infinitely far. As the pressure drops towards zero, the height of each step increases to infinity, and we never quite reach the top.

If this seems strange, consider it from another point of view. The molecules in a gas at a non-zero temperature have a non-zero speed; recall that their average kinetic energy is proportional to the temperature. If we reach the top of the atmosphere with a non-zero temperature, then the molecules near the top will still have non-zero speeds. They will therefore not stay below the point we have taken to be the top: some of them will shoot above it, so that there will be gas above the point we thought was the top. This is a contradiction. Therefore, for a gas in equilibrium, the top of the atmosphere is not only a place of zero pressure; it is also a place where the temperature must fall to zero. Although the isothermal atmosphere is easy to calculate, it is not a good approximation to whole planetary atmospheres.

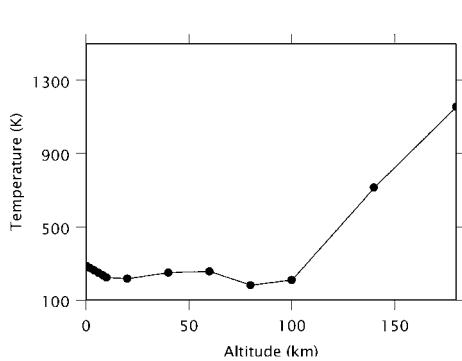


Figure 7.5. Measurements of temperature of the Earth's atmosphere (dots). The solid curve is simply a fit to the dots.

Before leaving this section, we pose a question we ignored earlier: why is it that the top of the atmosphere is marked by zero pressure? Why not by a finite positive pressure? (It can't be a negative pressure, because a gas cannot maintain a negative pressure. The physicists' term for negative pressure is **tension**.) If the atmosphere came to an end at a finite pressure, the gas just at the edge of the atmosphere would have a finite pressure below it, hence a finite force on it, with no force above it holding it down.

We could consider a very thin layer of such gas at the edge, with tiny mass, so that the finite pressure below it would blow it away. Such an atmosphere could not remain in equilibrium. Therefore, the pressure in an atmosphere that is in equilibrium must go to zero at the top, and indeed it must do so in such a way that a small sliver of air at the edge feels a small pressure below it just sufficient to balance its small weight.

The Earth's atmosphere

We can now look at the Earth's atmosphere in detail. The atmosphere is subject to many influences: physical forces from the Earth's rotation, heating from the Sun, and a huge number of chemical effects from chemical reactions and such processes as evaporation and precipitation. All of these affect the way the temperature behaves as one goes higher in altitude.

To try to understand this system in detail is a challenging research field. Our present understanding of some parts of the problem, such as the behavior of **ozone**, is seriously incomplete. One of the most important sets of data that researchers use in trying to understand the atmosphere is its *temperature profile*: the way the temperature behaves with altitude. If we use the measured values of the tempera-

In this section: we use a simple computer program to make an excellent numerical model of the Earth's atmosphere up to 250 km. Beyond that, the physics gets very complicated and the atmosphere cannot be represented as an ideal gas.

Investigation 7.3. Making an atmosphere

Our purpose here is to explore the ideas that go into our computer program *Atmosphere* that makes a realistic model for the atmosphere of the Earth. The program can be found on the website. It relies on the discussion in the text of this chapter up to and including the section on the constant-temperature atmosphere. We will find that this program will be easy to modify to construct models of the Sun and other stars in later chapters.

Our technique is to use the equation of hydrostatic equilibrium, Equation 7.1 on page 73, to move upwards in the atmosphere with constant steps in altitude h .

Suppose that we let the variable z stand for altitude. At the i^{th} step, we call the altitude z_i and the pressure and density p_i and ρ_i , respectively. Then at the next step the new altitude will be

$$z_{i+1} = z_i + h,$$

and Equation 7.1 on page 73 tells us that the new pressure will be

$$p_{i+1} = p_i - g\rho_i h. \quad (7.10)$$

To repeat this at the new height we need the new density ρ_{i+1} . We can get this from p_{i+1} if we know the temperature at this height, since then the ideal gas law, Equation 7.6 on page 78,

$$pV = NkT,$$

will give us the necessary information. Here is how to get the density from this.

The mass m of a parcel of air of a given volume V is the product of V and its density ρ . It is also equal to the number N of molecules it contains times the mass of each molecule. Since air is a mixture of gases, what we want is the average mass of a molecule. Now, molecules are basically composed of a few protons, an equal number of electrons, and a few neutrons. Moreover, the mass of a **neutron** is about equal to that of a proton, and is nearly two thousand times the mass of an electron. It follows that the mass of any molecule will essentially be the number of protons and neutrons times the mass of a proton m_p . The number of protons plus neutrons in a molecule is called its **molecular weight**. (For a single atom, scientists use the name **atomic weight** for the same thing.) The average mass of the molecules in a mixture of gases will be the average number of protons and neutrons times m_p . We shall call this average of molecular weights the **mean molecular weight** of the gas, and we use the symbol μ for it. Then we can write the **average** mass of a molecule of a gas as μm_p . For air, the mean molecular weight at sea level is $\mu = 29.0$. The two ways of calculating the mass of our parcel of air thus give

$$\rho V = N\mu m_p.$$

We can solve this for the ratio

$$\frac{N}{V} = \frac{\rho}{\mu m_p}. \quad (7.11)$$

From this and the ideal gas law we obtain

$$\rho = \left(\frac{\mu m_p}{k} \right) \frac{p}{T}. \quad (7.12)$$

This equation determines the density at each height, if we are given the pressure and temperature.

We have seen how we get the pressure by using the equation of hydrostatic equilibrium. How do we get the temperature?

As explained in the text, the temperature of the atmosphere is determined by a very complicated set of influences, and it would

be hopeless to try to model them in a simple computer program. Instead, we shall rely on the fact that the temperature of the atmosphere can be *measured* at different altitudes. These values are then put into the computer program, which uses them in the density calculation.

Our computer program, available on the website, handles this in a straightforward manner. Built into it are values for the fundamental constants, for μ , and for p_0 , the pressure at the bottom of the atmosphere. It also has the values of the altitude at which the temperature is measured, followed by the values of the temperature at those altitudes. These constitute a **temperature table**. As the comments in the program explain, the temperature at altitudes where it is not measured is approximated by taking the measured temperature at the two nearest altitudes and assuming it behaves as a straight line between them. Thus, if the temperature is known to have the values T_1 and T_2 at heights z_1 and z_2 , respectively, then its value at height z somewhere between is

$$T(z) = \frac{T_2 - T_1}{z_2 - z_1}(z - z_1) + T_1.$$

The program gets underway by computing a few useful numbers. In particular, it is useful to compute the so-called **scale-height** h_{scale} , which is a *rough* guide to the eventual height of the atmosphere. This is obtained by taking Δp in Equation 7.1 on page 73 to equal the pressure difference between the bottom and top of the atmosphere, and solving for h . Since at the top of the atmosphere $p = 0$, this gives

$$h_{\text{scale}} = p(0)/g\rho(0). \quad (7.13)$$

The program uses this as a guide for choosing the step size h . This is useful, for otherwise too small a value of h would waste computer time and too large an h would be inaccurate. The program ignores h_{scale} after this, so it does not assume that the atmosphere actually terminates there.

Then the program enters its main loop, stepping upwards in altitude until the pressure goes negative. There are some built-in safety measures to prevent the program taking too many steps.

There is one place where the program deliberately departs from realism, and that is at the top of the atmosphere. As explained in the text, the Earth's atmosphere becomes isothermal at high altitude, and so does not fit our notion of a finite atmosphere: it would in principle go on forever. In practice, the temperature is so high that the atmosphere is an **ionized** gas, and magnetic fields play an important role in what happens at these altitudes. The amount of material out there is so small, however, that we make little error at lower altitudes if we simply substitute an artificial cutoff at a high altitude. We do this by assuming that, above the highest altitude at which the temperature is supplied in the temperature table, the temperature is determined by the density by a relation of the form

$$T \propto p^{1/2}. \quad (7.14)$$

This is artificial, but it brings the atmosphere neatly to a termination. We will see in the next chapter that the relation between temperature and pressure inside stars is not unlike this equation.

The results of the calculation are plotted in Figure 7.4 on page 79. They are remarkably good, especially considering that our computational method is in essence very simple. We have not even introduced the predictor-corrector orbit program *Orbit*, and yet we still have been able to follow the density and pressure as they decrease by a factor of more than 10^5 .

Exercise 7.3.1: Finding values between measured points

Show that the equation used above for finding the temperature between measured points, $T(z) = (T_2 - T_1)(z - z_1)/(z_2 - z_1) + T_1$, does in fact describe a straight-line relationship between the height z and the temperature $T(z)$. Show that the line passes through the measured points (z_1, T_1) and (z_2, T_2) .

ture, as shown in Figure 7.5 on page 80, then we can construct an accurate model of the equilibrium atmosphere without understanding all the forces that shape the temperature profile. Data values from Figure 7.5 are used in the computer program on the website that we use to construct this model.

This temperature profile shows dramatic changes of temperature; the changes mark the boundaries of different layers of the atmosphere, where different physical processes take place. The temperature initially falls slowly with height, as anyone who has traveled in mountains would expect, until one reaches an altitude of about 15 km, where the stratosphere begins. In the stratosphere, the temperature gradually rises with height. This is mainly caused by the absorption of ultraviolet light from the Sun by the *ozone*. Then comes the mesosphere at about 50 km, where T falls again, until one reaches the thermosphere at about 100 km. From there the temperature rises dramatically and reaches a roughly constant value of 1500 K out to very great distances. So the outer regions of the atmosphere are in fact isothermal!

We have seen in the previous section that our simple atmospheres do not have a top if they are isothermal. In the case of the Earth, the outer regions are anything but simple! The gas there is ionized, magnetic fields are important, and the influence of the solar wind (see Chapter 8) begins to be important. Fortunately, the amount of mass in the outer reaches of the atmosphere is so small that whatever happens there has little effect on the structure of the lower atmosphere, and our computer program makes a very good model of this region, up to some 250 km (see Figure 7.4 on page 79).

Figure 7.4 shows that there are slight deviations between our model and the real atmosphere. These are mostly due to two effects we have left out: first, the composition of the atmosphere changes as one goes up; and second, the strength of the Earth's gravity decreases slowly as one goes higher. Despite these small differences, it is remarkable that we have been able to model the atmosphere so accurately with so little sophisticated mathematics.

If we want to go much beyond this, however, we are quickly humbled by the complexity of the physics. To explain the temperature profile would require an enormous computer program that contains not only the solar heating and the chemistry at different altitudes but also the dynamics of the atmosphere: convection of gas from one layer to the next, the influence of weather and storms, the many other time-dependent effects. To calculate the weather in any detail also requires complex programs, to take into account the variation of atmospheric properties from place to place, the effect of geography and variations in ocean temperature, the heating by the Sun, the transfers of energy and water between oceans and air, and so on. These areas are among the most active and challenging research areas in all of science today, and they make computing demands that exceed the capacity of the biggest available computers.

The atmospheres of other planets

In this section: temperature measurements on some other planets allow us to build models of their atmospheres too.

We can adapt the computer program to give us models of the atmospheres of other planets and indeed of some of the moons in the Solar System, just by replacing appropriate numbers. In Table 7.1 I have gathered the necessary data for three bodies: Venus, Mars, and Saturn's moon Titan. The output of the revised computer model for Venus is displayed in Figure 7.6.

Notice that all the temperature profiles include a temperature inversion: a place where the temperature begins rising again with altitude. This effect, which occurs because of the absorption of sunlight by the atmosphere, is seen in all Solar System atmospheres. Regarding Mars, the structure of its atmosphere is variable with time,

Venus				Mars				Titan				
$p_0 = 9.4 \times 10^6 \text{ Pa}$, $\mu = 43.2, g = 8.6 \text{ m s}^{-2}$				$p_0 = 730 \text{ Pa}$, $\mu = 43.5, g = 3.74 \text{ m s}^{-2}$				$p_0 = 1.6 \times 10^5 \text{ Pa}$, $\mu = 28, g = 1.44 \text{ m s}^{-2}$				
$h (\text{km})$	0	57	90	135	0	10	70	100	0	40	60	600
$T (\text{K})$	730	290	170	200	230	205	140	140	96	74	160	175

being strongly affected by dust storms, the seasons, and the latitude. Our model gives only an average structure. Titan, Saturn's largest satellite, is bigger than Mercury and retains a significant atmosphere. This is composed primarily of molecular nitrogen, but its history and evolution are shrouded in mystery. The temperature profile in Table 7.1 bears a striking resemblance to that of the Earth.

Quantum theory and absolute zero

To end this chapter, we turn to a matter of considerable importance to some topics we will discuss later in the book. I introduce it here in order to correct a misimpression that I deliberately allowed earlier, regarding what happens as the temperature of a gas is lowered to zero. According to Boltzmann, the motions of molecules will also go to zero as T goes to zero, so that a gas at absolute zero consists of molecules completely at rest (or at least it would if we could actually attain it). This turns out not to be quite what happens in the real world, because of the principles of the aspect of physics that we call **quantum theory**. We cannot go into quantum theory in any detail in this book, but we will use one of its most important principles in various places throughout the book: the *Heisenberg uncertainty principle*.

The brilliant German physicist Werner K Heisenberg (1901–1976) (see Figure 7.7 on the next page) was one of the founders of quantum theory. The principle that bears his name states that all measurable properties of any physical system come in certain pairs, which have a special relationship to one another: as one measures one member of the pair more accurately (say, with a better experiment), the other member inevitably becomes harder to measure accurately. Even given perfect measuring instruments, there is a minimum uncertainty in the measurement of the second member that is inversely proportional to the uncertainty in the measurement of the first. The constant of proportionality is known as *Planck's constant* h , and it has the value

$$h = 6.626 \times 10^{-34} \text{ kg m}^2 \text{ s}^{-1}.$$

We shall meet the man after whom the constant is named, Max Planck, in Chapter 10.

The value of h is so small that the uncertainties that are the subject of quantum theory rarely intrude into everyday measurements: our measuring instruments are typically too crude to be limited by quantum uncertainties. But it is not just actual measurements that are limited by these uncertainties: the limits apply to anything that *could in principle be measured*. This is how the uncertainty principle affects the behavior of matter.

A particularly important pair of measurable quantities consists of the position and the momentum of a particle. The

Table 7.1. Atmospheric data (altitude h and temperature T) for some Solar System bodies.

In this section: why atoms continue to vibrate even as as gas approaches absolute zero. These “zero-point” vibrations, a consequence of quantum theory, are responsible for many phenomena in the Universe that we will study later, involving neutron stars, black holes, and the Big Bang.

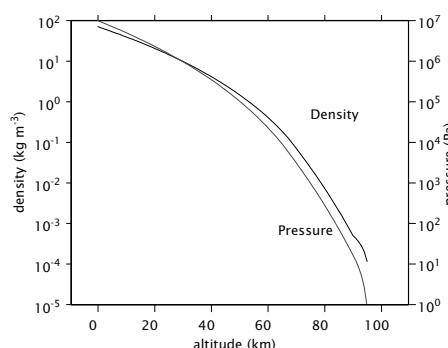


Figure 7.6. The atmosphere of Venus.



Figure 7.7. Werner Heisenberg formulated the theory now known as quantum mechanics when he was only 24 years old, during a holiday on an island in the North Sea where he had gone to escape hay fever. A year later a very different but equivalent formulation was achieved by the Austrian physicist Erwin Schrödinger (1887–1961). Heisenberg was a complex figure. Rather than leave Germany when Hitler began attacking Jewish physicists and even his own work on quantum mechanics, he remained and led Germany's unsuccessful program to build atomic weapons during the Second World War. After the war he continued to play a distinguished role in German physics. His wartime activities and his later attempts to portray himself as having been fundamentally opposed to creating nuclear weapons for Hitler have led to a continuing debate among historians and physicists about Heisenberg's character. His secret wartime visit to Niels Bohr (whom we will meet in Chapter 10) became the subject of the thought-provoking play Copenhagen, by Michael Frayn. Photograph courtesy Mary Evans Picture Library.

momentum is defined as the mass of the particle times its velocity, so the uncertainty here is basically between position and velocity. If the position is known (or could in principle be known) to great precision, then the particle's velocity would be very uncertain. This would have the consequence that a moment later, the position would be completely unknown: pinning the particle down at one moment forces it to squirt off in some completely random direction the next moment. For this reason, an accurate position measurement is not repeatable.

Now consider the gas at absolute zero. Boltzmann, who worked before the invention of quantum theory, would have expected that the molecules would be completely at rest. Thus, at least in principle, it would be possible to measure the position and (zero) velocity of each particle with arbitrary precision. Boltzmann would not have given this a second thought. But quantum theory forbids it.

Instead, what happens at absolute zero temperature is that each atom tries to reduce its velocity (hence its momentum) to as small a value as possible. This requires it to have as large an uncertainty in position as possible. If the gas is confined to a certain container, then the size of the container determines the maximum position uncertainty, and the minimum velocity uncertainty. Each particle retains a small average kinetic energy, whose size then depends on the volume of the container. In solids cooled to near zero, this is called the **zero-point energy** of vibration.

This illustrates the deep property of quantum theory, that the uncertainty is not just a value that is imprecisely known. The quantity that is uncertain can be thought of as undergoing random **quantum fluctuations** within this uncertainty. The energy of a cold particle is not just uncertain: the particle vibrates with this minimum energy even near the absolute zero of temperature.

In the Earth's atmosphere, the effect is completely negligible. But in some of the things we will study later, in systems as diverse as neutron stars, gravitational wave detectors, black holes, and the Universe itself, the uncertainty and its associated quantum fluctuations and quantum zero-point energy will be of critical importance.

Gravity in the Sun: keeping the heat on

We have seen how the Sun's gravity holds the planets in their orbits. The Sun's gravity also holds itself together. Like all stars, the Sun is a seething cauldron, its center a huge continuous hydrogen bomb trying to blow itself apart, restrained only by the immense force of its own gravity. In this chapter, we will see how the Sun has managed to maintain an impressively steady balance for billions of years. In the course of our study, we will learn about how light carries energy and we will build a computer model of the Sun.

Sunburn shows that light comes in packets, called photons

The Sun glows so brightly because it is hot. We can infer just how hot it is from its color. The color and temperature of the Sun are related to each other in just the same way as for hot objects on the Earth. For example, watch the burner of an electric stove as it gets hotter; it changes in color from black to red. It won't get any hotter than red-hot. But if you watch objects in a really hot fire, such as a blacksmith uses, you will see them change from red to a blueish white as they heat up.

As the temperature of an object increases, the radiation it emits moves toward shorter wavelengths, i.e. from red toward blue.

This change in color comes about in the following way. We saw in Chapter 7 that in hotter objects the molecules and atoms move faster. This means that when they collide and emit radiation, the radiation usually has higher energy. Now, it is a remarkable fact, which we will explore here, that higher-energy radiation has shorter wavelength: blue light is more energetic than red. It follows from these two observations that hotter objects tend to be bluer.

That light carries energy is obvious to everyone: the warmth of sunlight is caused by the conversion of the energy carried by the light into thermal energy (random kinetic energy) in our bodies. The fact that light of a certain *color* carries a *specific amount of energy* is a deeper property of physics, but it can be illustrated with an equally commonplace event: getting sunburned.

On a clear hot day, if you have sensitive skin, it does not take long to get a good red sunburn. But if you apply a blocking sunscreen lotion, you can remain in the same sunlight for hours without a burn. The lotion acts like a "filter" that prevents light of wavelength shorter than a certain ultraviolet wavelength from reaching your skin. No matter how much light of other colors reaches the skin, no matter how much energy in total the sunlight transfers to your skin, if it does not have a short enough wavelength it will not do the damage. There is clearly something different about the longer wavelengths of light. We will see that the difference is that the longer wavelengths of light do not carry enough energy to set off the chemical reactions in the skin that lead to sunburn.

The relation between the energy and the wavelength of electromagnetic radiation was discovered by Einstein. It was part of his explanation of the **photoelectric**

In this chapter: we learn how the Sun holds itself up. The key is another discovery of Einstein, that light actually comes in packets called photons. These form a gas that helps support the Sun. Photons move randomly in the Sun, taking millions of years to get out. We compute the structure of the Sun, and learn why stars and planets are round, while asteroids and comets are lumpy. Finally we study the vibrations of the Sun, which reveal the details of the Sun's interior to astronomers.

In this section: to understand stars, and in particular the Sun, we first learn about photons: packets of light whose energy is proportional to their frequency. The simple phenomenon of sunburn illustrates the way photons behave. The idea of a photon was first introduced by Einstein.

>The image beneath the text on this page is a picture of the Sun taken by the SOHO spacecraft on 14 September 1999, through a special filter. It shows a *superprominence*, the large loop of hot gas streaming out of the Sun. When such a prominence moves towards the Earth it can disrupt communication and electricity supplies, and cause aurora. The Sun is a turbulent, violent ball of gas that is only kept together by the strong force of its self-gravity. Image courtesy NASA/ESA.

effect, which is a metallic version of sunburn. It had been observed that light falling on certain metals can eject electrons, but only if the light has a short enough wavelength. This threshold wavelength depended upon the metal. As the wavelength of the light decreased further, the electrons came out with more and more kinetic energy.

Einstein proposed that light actually comes in discrete packets, which we now call **photons** or **quanta**. Each photon carries a fixed amount of energy that can be transferred to an electron or other particle if the photon collides with it. This energy can then be converted into kinetic energy of the electron. Einstein suggested that the energy of a photon is determined entirely by its wavelength: the shorter the wavelength, the more energy. He then proposed that each metal has what is effectively an “escape speed” caused by the attraction of molecular forces inside the metal, so that if the kinetic energy given to an electron by a photon were too small, it would not attain this speed and would therefore not be ejected. Once the wavelength of the photon was short enough to give the electron its escape speed, the electron would use up a certain amount of its kinetic energy escaping, and the rest would turn up as kinetic energy of the ejected electron. This is analogous to what happens when spacecraft escape from the Earth.

►Physicists call the minimum energy for escape the *work function* of the metal.

This neatly explained all the experiments on the photoelectric effect, but it was nevertheless a revolutionary step in physics. Physicists had been used to thinking of light as a wave. A water wave’s energy depends on its height, not its wavelength: we avoid swimming in the sea if the waves are large, not if they have very short spacing! The idea that light waves carried energy in discrete amounts, which depended on the wavelength, meant that scientists had to start thinking about light as if it were a particle. This took some getting used to.

But the experimental evidence in favor of Einstein’s proposal is overwhelming, and this so-called wave–particle duality of light is something that modern physics has come to embrace, even if it is a little hard to visualize in concrete terms. It is a fundamental aspect of quantum theory. Light behaves like a wave in some respects, for example when it refracts or interferes, and like a particle in other respects, such as by carrying fixed amounts of energy.

We shall more often refer to photons in the rest of this book than to light waves. Photons make a host of astronomical facts easier to understand.

The relation between wavelength and energy that Einstein proposed is remarkable because Einstein did not need to introduce a new constant of Nature to make the theory fit the observations: he only needed to use ones that were already known to be important for the physics of light: the speed of light, c , and Planck’s constant h . Planck’s constant had only recently been introduced by Max Planck to describe the spectrum of the radiation emitted by hot bodies. We have already encountered it in Chapter 7, where we saw how it plays a fundamental role in the uncertainty principle. We shall introduce its importance for light here, but defer a discussion of Planck’s original use for it until we study the colors of stars in general in Chapter 10.

►The Greek letter λ , called *lambda* and pronounced “lam-da”, is standard physics notation for the wavelength of a wave.

Einstein showed that the energy carried by a photon of wavelength λ is inversely proportional to λ , the constant of proportionality being h times the speed of light c :

$$\text{energy } E \text{ of a photon} = hc/\lambda. \quad (8.1)$$

This relation is described more fully in Investigation 8.1.

Investigation 8.1. The colors of energy

Here we learn how to find the energy carried by a photon of a given color. Einstein's postulate for the photoelectric effect led, with other developments, to the quantum theory. In quantum theory, light (or any other electromagnetic radiation) is really composed of **photons**, which can be thought of as little packets of energy. The amount of energy E carried by each photon packet is directly proportional to the *frequency* of the light, and the proportionality constant is h , *Planck's constant*, which we met in Chapter 7:

$$E = hf, \quad (8.2)$$

where f is the frequency of the light (measured in units of cycles per second, which scientists call Hertz, denoted Hz). Notice that this equation is in fact a further illustration of the close relationship between energy and time that we first met in Chapter 6.

Because we often think in terms of the *wavelength* λ of light rather than its frequency f , we shall convert this equation using the relation between wavelength and frequency for a wave whose wave speed is c (which in our case is the speed of light):

$$f = c/\lambda, \quad (8.3)$$

where the speed of light has the value $c = 2.998 \times 10^8 \text{ m s}^{-1}$. Then we find that the energy of a photon is

$$E = hc/\lambda. \quad (8.4)$$

Exercise 8.1.1: Frequency of light

Find the frequency (in Hz) of light whose wavelength is $0.5 \mu\text{m}$.

Exercise 8.1.2: Photons from a light-bulb

Show that a 100 W light bulb (which emits 100 J of energy each second) must be giving off something like 10^{21} photons per second.

Exercise 8.1.3: Sunburn

The DNA molecules that carry genetic information in the nuclei of living cells are very sensitive to light with a wavelength of $0.26 \mu\text{m}$, which breaks up DNA molecules. Deduce from this the binding energy of the chemical bonds within DNA. Ultraviolet light of wavelength $0.28 \mu\text{m}$ is the most effective for inducing sunburn. What is the threshold energy required to stimulate the chemical reactions that lead to sunburn?

Exercise 8.1.4: Gamma-rays

When some elementary particles decay, they give off so-called **gamma-rays**, which are really high-energy photons. A typical energy released in this way is 10^{-12} J . What is the wavelength of such a gamma-ray? What is its frequency?

A gas made of photons

If photons behave like particles, colliding with electrons and exchanging energy with them, then there can be circumstances in which it would make sense to speak of a *photon gas* mixed with an ordinary gas of electrons and ions, in which collisions between photons and gas particles would be as common as between the gas particles themselves. This happens inside stars, where the density of gas is so great that, as we explain later in this chapter, the photons bounce off atoms of the gas a fantastic number of times before they reach the surface. At each bounce they exchange energy with the gas.

This has two important effects. First, the collisions with photons exert *pressure* on the gas particles; and second, the exchange of energy with gas particles means that the photons in any part of the Sun come into *equilibrium* with the gas: the typical energy of a photon is roughly the same as that of a gas particle. Thus, the photon gas really behaves like a gas: it has a temperature and a pressure.

Now, the energy of a gas particle is determined by the temperature of the gas, and this in turn must equal the energy of a typical photon if collisions occur often enough to insure that energy is frequently exchanged between particles.

There is thus a characteristic photon wavelength associated with any

Visible light has a wavelength in the range $0.4\text{--}0.7 \mu\text{m}$. (One μm is 10^{-6} m , and is sometimes called a **micron**. Readers who are used to old-style units may prefer **Ångströms**; one micron is 10^4 Å .)

If we insert the values of h and c into the previous equation and then multiply it by $1 \mu\text{m}/1 \mu\text{m}$ (which equals 1, of course), we get

$$E = \left(\frac{1.986 \times 10^{-25} \text{ J m}}{\lambda} \right) \left(\frac{10^{-6} \text{ m}}{10^{-6} \text{ m}} \right) \\ \approx 2 \times 10^{-19} \left(\frac{10^{-6} \text{ m}}{\lambda} \right) \text{ J}. \quad (8.5)$$

This is a handy way of writing Equation 8.4 in a way that shows the scale of energies involved, and allows one to do the arithmetic more easily in one's head. A $1 \mu\text{m}$ photon (infrared light) carries about $2 \times 10^{-19} \text{ J}$ of energy. A green photon with a wavelength of $0.5 \mu\text{m}$ has twice this energy.

These energies are very small in everyday terms, as Exercise 8.1.2 shows. From the result of that exercise, it is not surprising that the eye is unaware of the discrete nature of the packets of energy that keep striking it: so many arrive per second that they merge into a continuous stream of energy. But these packets, or quanta, of energy do play an important role in a huge variety of situations, from the workings of individual atoms to the structure of stars.

In this section: radiation can form a gas of its own, which provides some of the pressure that holds up stars. We estimate the temperature of the radiation from the Sun from its color. We also introduce a new unit for energy, the electron volt.



Figure 8.1. Albert Einstein in 1905, when he was still working at the patent office in Bern, Switzerland. He revolutionized research in three different fields with his scientific papers that year. Photo courtesy ETH Library Picture Archive, Zürich.

►The reason for the name *electron volt* is that 1 eV is the energy acquired by an electron when it is accelerated by an electric field corresponding to a difference in potential of 1 V. Electrons inside television monitors are accelerated by a difference of several thousand volts before they are directed at the screen.

In this section: Einstein published five extraordinary papers in a single year, 1905, making breakthroughs in three different problems.

►This was the first discussion in physics of the *random walk*, to which we will return later in this chapter when we discuss the diffusion of light through the Sun.

temperature. The wavelength of the photons emitted by the Sun is an indication of the temperature of the gas near the Sun's surface.

Now we can come back to sunburn and use it to estimate the Sun's temperature. The fact that the Sun emits significant amounts of ultraviolet light means that we will get what physicists call an "order-of-magnitude" answer if we just set the thermal kinetic energy of particles in the Sun, $3kT/2$, equal to the energy of an ultraviolet photon, which we computed in Exercise 8.1.3 on the previous page to be about 7×10^{-19} J. Solving for T gives $T = 34\,000$ K. We should expect this to overestimate the temperature of the Sun, perhaps by as much as a factor of 5 or 10, since most of the Sun's light comes out in the visible region, not the ultraviolet. But at least it tells us that the surface of the Sun is at least several thousands of degrees, but not as high as several million. Often in physics such order-of-magnitude estimates are all one needs to get a reasonable understanding of a physical phenomenon.

Here is the appropriate place to introduce a new unit of measure for energy, one that is better suited to the tiny energies of photons and atomic interactions than the joule: it is not very convenient to keep writing numbers like 10^{-19} J. Physicists have introduced the unit called the **electron volt**, abbreviated eV, which is equal to

$$1 \text{ eV} = 1.602 \times 10^{-19} \text{ J}. \quad (8.6)$$

The energy carried by one-micron (infrared) light is thus about 1.24 eV.

We do a better job of estimating the Sun's temperature in Investigation 8.2, by using the wavelength at which the Sun is brightest, and we improve on this method even further in the next chapter. The temperature of the surface of the Sun is actually about 5800 K.

At this temperature, ordinary materials cannot exist in the solid or liquid state: the Sun is a ball of hot gas consisting of electrons and ions, called a **plasma**. It is important to remember that the temperature we measure from the color of the Sun is only its *surface* temperature, since the light we see comes only from a thin surface layer called the **photosphere**. Inside the Sun temperatures rise sharply, to around 10^7 K in the center. We shall see why later in this chapter. We don't see the light from this region directly because of all the collisions that photons undergo inside the Sun. We only see the photons that have finally escaped after their last collision.

Einstein in 1905

Einstein's paper on the photoelectric effect was one of *five* landmark papers that he published in one extraordinary year, 1905. Besides introducing the quantum nature of light in order to explain the photoelectric effect, his other papers were equally revolutionary. Two of them established the theory of special relativity, to which we will return in Chapter 15. The other two explained the so-called **Brownian motion**, in which microscopic specks of dust floating on the surface of water had been observed to execute completely random motions. Einstein showed that huge numbers of random collisions with molecules, each making an unobservably tiny change in the motion of the dust, would add up to the observed motions. The Brownian-motion papers helped to establish the correctness of Boltzmann's atomic theory, and Einstein was even able to calculate for the first time the average masses of the molecules.

Einstein's paper on the photoelectric effect was one of the papers that founded quantum theory. The fact that Einstein founded three fields of physics in a single year (while holding down a full-time job in the patent office in Bern, Switzerland) is as impressive an indication of his genius as is his monumental work on general relativity ten years later, which occupies the second half of this book.

Investigation 8.2. How hot is the Sun?

Here we want to infer the temperature of the Sun from the wavelength of the radiation it emits. We can rewrite Equation 8.5 on page 87 in terms of electron volts to give

$$E = 1.24 \left(\frac{1 \text{ } \mu\text{m}}{\lambda} \right) \text{ eV}. \quad (8.7)$$

As we mentioned in the text, there is a characteristic photon wavelength associated with any temperature. This is given by setting E in Equation 8.7 equal to $\frac{3}{2}kT$ from Equation 7.3 on page 76, where k

is Boltzmann's constant. The result is that

$$\lambda \approx \left(\frac{9600 \text{ K}}{T} \right) \mu\text{m}. \quad (8.8)$$

In the center of the Sun, where the temperature reaches 10^7 K , photons have typical equilibrium wavelengths of $0.001 \mu\text{m}$ and energies exceeding 1000 eV . Throughout the Sun, the energies of photons are enough to ionize hydrogen, and to strip electrons from other atoms too. Such a gas of electrons and ions is called a *plasma*.

Exercise 8.2.1: Temperature of the Sun

If one analyzes the colors of the Sun, one finds that the greatest amount of light is emitted in the *blue-green* region of the spectrum, around $0.5 \mu\text{m}$. Show that this gives an *estimate* of the Sun's temperature of $T = 19000 \text{ K}$. This is closer to the real temperature (5600 K) than our estimate in the text, but we will get a much better estimate by refining this technique in the next chapter. (The eye sees the Sun as yellow, not blue-green, partly because it has greater sensitivity to yellow light and partly because blue light is scattered by the atmosphere.)

Einstein was also the right man at the right time. Physicists were just learning how to probe the world of atoms and particles, and this world did not behave the way they expected, based on their experience with **macroscopic** objects. Einstein had an extraordinary ability to free his mind from prejudices and begin thinking in the new ways that atomic physics demanded. And not just to think, but to calculate, to make predictions that could be tested by experiment.

Interestingly, when Einstein received the 1921 Nobel Prize for physics, it was for the photoelectric effect. His work on relativity was explicitly excluded, since in the eyes of the awarding committee, it had not yet been sufficiently confirmed.

Gravity keeps the Sun round

We know the mass ($1.99 \times 10^{30} \text{ kg}$) and radius ($6.96 \times 10^8 \text{ m}$) of the Sun, and from them we can work out that the mean density (mass divided by volume) of the Sun is about 1400 kg m^{-3} , or 1.4 times the density of water. To compress a gas whose interior temperature is several million degrees to beyond the density of water requires a great deal of force.

What is this force in the Sun? The answer can only be gravity, the gravitational attraction of one part of the Sun for another. This mutual attraction would, if unresisted, simply pull the material of the Sun inward towards a single point. The resistance to this collapse is provided mainly by gas pressure, and secondarily by the pressure provided by all the photons that are produced in the center of the Sun and gradually make their way outwards, scattering off electrons and nuclei in the Sun countless times as they go. The Sun exists in a state of balance between the outward push of gas and radiation pressure (the pressure of the photon gas) and the inward pull of gravity.

The photons produced in the center come from **nuclear reactions**, which are processes that change nuclei of some atoms into other nuclei. These reactions release a great deal of energy, and are the chief source of the energy that makes the Sun (and all other stars) shine. The energy from these reactions leaves the Sun in two main forms: as photons and as **neutrinos**, which are very light particles produced in many nuclear reactions. We will look at how nuclear reactions work in Chapter 11. For now, we just assume that there is an energy source in a small region around the center of the Sun.

The shape of the Sun is also determined by gravity. Gravity is, like pressure, an *isotropic* force, that is a force that has no preferred direction: the gravitational attraction exerted by any particle is the same in all directions. As long as the Sun is

In this section: gravity singles out no special direction, nor does pressure, so stars and other large bodies are basically round. However, smaller bodies can be irregular in shape if chemical forces are significant. We calculate that any body with more mass than $1/1000^{\text{th}}$ of the mass of the Earth should be round: gravity should dominate chemistry. This fits well with observations in the Solar System.

a balance between pressure and gravity, it can't help but form a ball that is round. If there were corners or other special places on its surface, then if we stood at the center of the Sun and looked outwards, there would be some directions different from other directions. Since there is nothing in gravity or pressure to single out these directions, they cannot exist: the Sun should be a sphere.

We have left out of this discussion three extra influences that *can* single out directions: rotation, a magnetic field, and the presence of a nearby gravitating body, such as a companion star in a binary system. We will discuss rotation below, since it does affect the Sun. Many stars have magnetic fields similar to that of the Earth, with a North and South magnetic pole. Pulsars, which we shall study in Chapter 20, have fields an incredible 10^{12} times stronger than the Earth's. If the field is strong enough, it can cause the star to have a distorted shape, particularly near the poles. The Sun's field is not that strong. Even the superprominence illustrated in the figure on page 85 contains a negligible amount of mass. Jupiter acts like a companion "star", distorting the shape of the Sun by tidal effects (see Chapter 5), but the effect is too small to measure. We will return to a more detailed discussion of the distorting effects of companions when we meet binary stars in Chapter 13. All in all, the Sun has no choice but to be spherical.

The arguments of the last paragraphs apply in fact to any astronomical bodies for which gas pressure and gravity are the main forces. But there are many astronomical bodies that are not round, because other effects dominate. Dust grains are whisker-shaped because of chemical forces; **asteroids** are irregular because the chemical forces that shape their rocks are as important as gravity; and on a very large scale, galaxies can be disk-shaped or cigar-shaped because the motions of large numbers of individual stars define their outlines.

In the case of asteroids, we can combine Boltzmann's understanding of the kinetic energy of a molecule (Chapter 7) with what we learned in Chapter 6 about escape speeds to answer the following elementary question.

Why are planets and moons round, while ordinary rocks and even asteroids and the cores of comets have corners?

Figure 8.2. Phobos is one of the two moons of Mars, and has a mass smaller than the number we calculate in Investigation 8.3 to be the minimum mass for a rocky body to be forced by gravity to be round. Its irregular shape is consistent with our calculation.

Image courtesy NASA.



The answer has to do with melting. If, when a body was formed in empty space, temperatures got high enough to melt it, then gravity would, as we have just argued, make it spherical, as long as rotation, magnetic fields, and tidal effects were not too important. Since the atoms and molecules that form planets heat one another by colliding as they fall together, their kinetic energy when they collide must be comparable to their gravitational potential energy. Given a molecule of mass m , falling onto a planetary body of mass

M and radius R , its kinetic energy when it arrives will be something like GMm/R . The collisions randomize the direction of this energy, turning it into heat. The temperature, according to Boltzmann, will be given by setting $3kT/2$ equal to this energy. In Investigation 8.3 we put these expressions together and find that a rocky body in our Solar System should be round if its mass exceeds 3×10^{21} kg. This number depends on some assumptions, especially that the body is composed of silicate

Investigation 8.3. Why the Moon is round

We do a little algebra here to find the minimum mass M a body would need to have in order to melt as it forms. Once molten, gravity will shape it into a sphere. But if it does not melt, then it can have any irregular shape.

A molecule of mass m falling onto the body will have a kinetic energy approximately equal to GMm/R , where R is the size of the body (its radius, if it is spherical). This is an approximation, and in fact the energy could be more, but we are only interested in a rough answer. The collision transforms this into random kinetic energy, or heat, with a temperature T given by Boltzmann's relation

$$\frac{3}{2}kT \approx \frac{GMm}{R}.$$

We want to find the mass M required to make T high enough to melt the material. So we will assume that we know T , and set it equal to the melting point of rocks when we start doing numbers below. We shall also take m to be the mass of a typical molecule in a rock crystal when we do the numerical calculation. So we would like to solve this equation for M in terms of T and m , but we don't yet know R , the size of the body. To get a sensible value for R let us assume that we know the density ρ of the body, which we will take later to be the density of rock. Knowing ρ gives us a further approximation (again assuming a roughly spherical body)

$$\rho \approx M / \left(\frac{4}{3}\pi R^3 \right),$$

which can be solved for R to give

$$R \approx M^{1/3} \rho^{-1/3}.$$

Exercise 8.3.1: Rounding off the Moon

Do the algebra that leads to Equation 8.9 from the two equations that precede it. Then put the given numbers into the formula to arrive at Equation 8.10.

From now on I will ignore factors like the $4\pi/3$ in the density equation, since our answers are only going to be rough order-of-magnitude approximations anyway. If we put this into our first equation and solve for M we find

$$M \approx \left(\frac{kT}{Gm} \right)^{3/2} \rho^{-1/2}, \quad (8.9)$$

again dropping simple numerical factors.

Now, Solar System bodies typically have the density of rocks, about $\rho = 6000 \text{ kg m}^{-3}$. Let us take the molecule to be SiO_2 , the main constituent of sand. The silicon **nucleus** contains 14 protons and 14 neutrons, and each oxygen nucleus contains 8 protons and 8 neutrons. Altogether, there are 60 protons and neutrons in one molecule. The mass of the molecule is about 60 times the mass m_p of a proton. (We neglect the small mass difference between protons and neutrons, and we neglect the mass of the electrons, which are a fraction of a percent of the mass of the nuclear particles.) Looking up the mass m_p in the Appendix, we find that the mass of the molecule is $m = 1 \times 10^{-25} \text{ kg}$. Finally, the melting temperature of silicon dioxide is about 2000 K. Putting all these into Equation 8.9, we find the minimum mass of a round body in the Solar System to be about

$$M \approx 3 \times 10^{21} \text{ kg}. \quad (8.10)$$

This is about 5% of the mass of the Moon, and much larger than the mass of any known asteroid or comet.

rocks. For icy bodies, the mass would be a bit smaller. We should treat this as an order-of-magnitude estimate of the smallest mass of a round astronomical body.

For comparison, the mass of the Moon is $7.3 \times 10^{22} \text{ kg}$, so its round shape is no surprise. Our minimum "round" mass is much larger than the mass of any known asteroid or comet, so we should expect asteroids and comets to have irregular shapes, as indeed they all do. In fact, many planetary moons in the Solar System are of smaller mass. For example, Mars' moon Phobos has a mass of 10^{16} kg and is very irregular, as Figure 8.2 shows. In fact, the largest irregular body in the Solar System is Saturn's moon Hyperion, whose mass is $1.8 \times 10^{19} \text{ kg}$. So our rough calculation is not bad.

The Sun is one big atmosphere

Because the Sun is a balance between gravity and pressure, it is like one giant atmosphere. All the discussion of Chapter 7 can be directly applied here to help us understand the Sun's structure. We will extend the computer program of that chapter to help us make a numerical model of the Sun, and in the next chapter we will apply it to building other stars as well.

What changes do we need to make to apply our atmosphere program to the Sun? One obvious one is that the atmosphere program assumed we were dealing with the gravity of the Earth, not of the Sun. Changing this is not just a matter of changing the value of g , the acceleration due to gravity. For an atmosphere, which is a thin layer sitting on top of a big planet, one can assume without losing too much accuracy that the acceleration due to gravity is the same everywhere in the atmosphere. This assumption does not work for the Sun. For example, at the center of the Sun the acceleration must be zero, since particles are being pulled by the different parts of

In this section: the structure of the Sun is described by the same basic equations as we used to determine the structure of the Earth's atmosphere.

Investigation 8.4. How the gas in the Sun behaves

As mentioned in the text, we do not have direct measurements of temperature inside the Sun. We instead assume that the physics inside the Sun can be summarized by a relatively simple *equation of state*: a relation between pressure, temperature, and density. We shall use what physicists call a **power-law** relationship between density and pressure, that is a relationship where one variable is proportional to the other raised to a constant power (exponent). The usual way physicists write this is:

$$p = C\rho^\gamma. \quad (8.11)$$

Astrophysicists call this a *polytropic* equation of state, which is another word for “power-law”, but they adopt a somewhat strange way of writing the exponent. They use a *polytropic index* n in place of the polytropic exponent γ , defined by

$$n = \frac{1}{\gamma - 1}, \quad \text{or} \quad \gamma = 1 + \frac{1}{n}. \quad (8.12)$$

A star with a polytropic equation of state is called a polytrope.

Given the ideal gas law (Equation 7.12 on page 81) for a gas with mean molecular weight μ , we can solve it for the pressure to obtain

$$p = \frac{k}{\mu m_p} \rho T. \quad (8.13)$$

It follows that the temperature can be expressed in terms of the pressure by the power-law given in Equation 8.15,

$$T = A p^\beta,$$

with the constants

$$\beta = 1 - \frac{1}{n+1} = \frac{1}{n+1}, \quad \text{and} \quad A = \frac{\mu m_p}{k} C^{1/\gamma}, \quad (8.14)$$

where C is the proportionality constant in Equation 8.11. The constant C can be determined if the pressure and temperature are given in one place, say at the center of the Sun.

Assuming the Sun to be a polytrope is in fact not as arbitrary as it might seem. In regions that are dominated by convection of heat from the interior of the Sun, and in regions where most of the pressure is provided by the radiation making its way outwards through the Sun, the equation of state does indeed follow such power-laws. The physics hidden behind this statement is a little beyond our scope here. We shall simply adopt the polytropic equation of state and look at the models it produces.

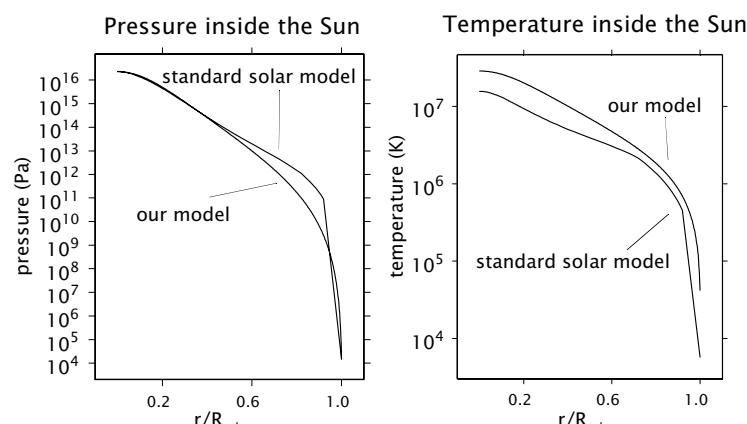
the Sun equally in all directions.

The acceleration due to gravity inside the Sun is not hard to compute, however. Recall that in Chapter 4 we saw that inside a spherical shell, the force of gravity due to the shell is zero, whereas outside it the force is the same as if all the mass were concentrated at the central point. If we consider a point inside the Sun, then if we draw a sphere about the Sun’s center through the point in question, the sphere divides the Sun into an “inside” and an “outside”. The material outside the sphere does not contribute to the gravity, while that inside acts from the center. The acceleration due to gravity at the point is, therefore, just the acceleration due to that part of the mass of the Sun that is within the radius in question. This means that the computer has to keep track of how the mass of the Sun is increasing as we go out in radius, but computers are good at doing such things.

The other part of the program that needs to be modified is the computation of the temperature. This is not quite as easy to handle as gravity, because it requires a discussion of gases that goes beyond our treatment in Chapter 7. This is the subject of the next section.

Figure 8.3. Comparing our solar model with the Standard Solar Model. Our computer model assumes that the relation between pressure p and density ρ is

$p \propto \rho^{1.357}$ everywhere, and then adjusts the constant of proportionality and the central pressure to set the model’s total mass and radius to those of the Sun. Of course, this relation is only an approximation to the real physics inside the Sun, so our model cannot describe all the details of the Standard Model. In particular, it overestimates the temperature everywhere.



The Standard Model of the Sun

The biggest difference between modeling the Sun and modeling the Earth's atmosphere is that the interior of the Sun is not directly observable: we can't send a spacecraft into it to measure the temperature down in the center! In Chapter 7 we were able simply to specify the temperature of the Earth's atmosphere as a function of height, allowing direct observation to replace detailed modeling of the chemistry and physics that lead to the specific temperatures observed. For the Sun, we have no choice: we must put in all the physics to make the best model. When all this is done – including the nuclear energy generation in the center, the pressure of radiation flowing outwards through the gas, and the convective motions of the gas as it is heated from below – and when all the unknown variables are adjusted so that the model matches whatever we can observe – such as the radius, mass, composition, age, neutrino radiation, and total energy output of the Sun – the resulting model is called "the Standard Model" of the Sun.

We can't attempt here to recreate the Standard Solar Model: many of the best astrophysicists in the world spend their working lives trying to make good models of the Sun and other stars! Instead, we will employ some approximations that are well grounded both in the science of gases – thermodynamics – and in experience in astrophysics. We have seen in Investigation 8.2 on page 89 that the trapping of photons in the Sun brings the photons into temperature equilibrium with the gas. It is therefore possible as a first approximation to ignore the *flow* of energy outwards and to treat the Sun as a static ball of gas and radiation. Then the next simplifying approximation is to assume that the pressure and temperature are related to one another by the following simple expression, called a *power-law*:

$$T = A p^\beta, \quad (8.15)$$

where A and β are constants.

Although this may look like a great over-simplification, we will see that if we choose the power β appropriately, we can make a fair approximation to the Standard Solar Model. A star that has such a power-law equation of state relating pressure, density, and temperature is called a **polytrope**. We explore these ideas further in Investigation 8.4.

The structure of the Sun

In Investigation 8.5 on the next page we assemble the various elements of our discussion above into an outline of the computer program we use to construct our model of the Sun. In this section, we can take a look at the output from this program to see what we can learn about the Sun itself. We display the output in the form of the two graphs in Figure 8.3. These graphs show the pressure and temperature of both our model and the Standard Model, as a function of radius within the Sun. The radius is expressed as a fraction of the full solar radius R_\odot , which is $R_\odot = 6.96 \times 10^8$ m.

In these graphs we compare the output of our computer program with one of the most recent versions of the Standard Model. To arrive at the model displayed in this figure, I have experimented with a few choices of the value of the index n . For each n , I found the values of the central pressure and temperature that gave the model the same total radius and mass as the Sun. The value of n that seemed to give the best overall approximation to the Standard Model was $n = 2.8$, whose model is shown in Figure 8.3. (Pressure is shown in the first graph in Figure 8.3, but density is not displayed.) The appropriate values of central pressure and temperature are given in the computer program Star on the website.

In this section: we discuss how to modify the computer program we used for the Earth's atmosphere to make a model of the Sun. When astrophysicists put in all the physics that they believe to be relevant in the Sun, they obtain what is called the Standard Model, their best guess about what the Sun is like deep inside.

►The Sun's age must be greater than that of the oldest rocks on the Earth, but presumably not much greater.

►A power-law is just a relationship between two quantities in which one of them is proportional to the other raised to a fixed power.

►Of course, all the complicated physics is hidden away in determining the right values of β and the constant of proportionality. We will simply choose the values that give us the best approximation.

In this section: we discuss the features of our computer model of the Sun. Although we have left out much of the physics, the model is a remarkably good approximation to the Standard Model.

Investigation 8.5. A computer model of the Sun

The changes that we need to make to our planetary atmospheres computer program Atmosphere on the website to adapt it to the Sun are relatively minor, and they make the program considerably *simpler*. The result will be a new computer program called Star.

The first change is to recognize that the acceleration due to gravity changes as one goes outwards through the star. As we noted in the text, the local gravity at any distance r from the center of the Sun depends only on the mass inside the sphere of radius r . We therefore define a variable called $m(r)$ (the array M in the program) whose value is the mass inside a sphere of radius r . If we move outwards from the sphere of radius r to one of radius $r + h$, with h very small compared to r , then the difference between $m(r)$ and $m(r + h)$ will be the mass inside the thin shell of thickness h . If the shell is thin enough, then the density inside it will be essentially the same everywhere, and equal to $\rho(r)$. The volume of such a shell is its thickness h times its area $4\pi r^2$, so we have the new equation for the increase of mass as one goes outwards:

$$m(r + h) = m(r) + 4\pi r^2 \rho(r) h. \quad (8.16)$$

Thus, as we go outwards, we need not only to decrease the pressure according to the equation of hydrostatic equilibrium, Equation 7.1 on page 73, but also to increase the mass at each step. This leads to other minor changes in the program, such as (1) the fact that in place of the acceleration of gravity g in the pressure equation we have to use Newton's law,

$$g = \frac{Gm(r)}{r^2}, \quad (8.17)$$

and (2) in the computer program the variable G no longer stands for g but instead for Newton's constant G .

The second change is the treatment of temperature. Since we adopt the polytropic law in Equation 8.15 on the previous page, we can get rid of all the steps in the program Atmosphere that had to do with reading in the temperature profile and using it to calculate the temperature at any height. Instead, once the pressure is known at any radius r , the temperature and density can be calculated immediately from it.

These are the only two changes in the program that come from the physics. We need also, however, to consider a technical point about how to get the computer to solve the equations. This is because there is one danger lurking in Equation 8.17. Since the radius of the Sun starts out at zero at the center, a careless programmer could wind up asking the computer to divide by zero to calculate the

local acceleration of gravity. Computers don't like to do this, and they usually crash, stopping the program with an error.

First of all, we must be sure that there is no real problem with Equation 8.17 at the center of the Sun: $m(r)$ is zero there too, so we must see what happens near the center, as r gets smaller. Mathematically, we say we are looking at the *limit* $r \rightarrow 0$ to see whether the ratio $m(r)/r^2$ is in fact well-behaved.

The density reaches a maximum at the center: let us call it ρ_c (the variable rhoC in the program). Consider a tiny sphere of radius r about the origin. Within it, the density will not change much from place to place, so its mass will be the density times the volume of the sphere,

$$m(r) \approx \frac{4}{3}\pi r^3 \rho_c \quad \text{for small enough } r.$$

Then the acceleration of gravity at such a small radius r is

$$g(r) \approx \frac{4}{3}\pi G \rho_c r.$$

In the limit as r gets smaller and smaller, this equation becomes exact (the approximation sign \approx is replaced by equality), and we have

$$\lim_{r \rightarrow 0} g(r) = 0.$$

It should not be surprising that the acceleration of gravity at the center is zero: the Sun's gravity is pulling on the center equally in all directions, thus cancelling itself out and leaving no net pull at all.

The problem is therefore only to avoid asking the computer to perform a division at zero. We take the easiest way around this: we start the computer at radius h , the first step away from the center. We set the values at the center to the obvious ones: $m = 0$, $p = p_c$ (given as initial data), $T = T_c$. This allows one to find the constant C and therefore to find $\rho = \rho_c$ from the polytropic equation Equation 8.11 on page 92. Then at $r = h$ we set $m = \frac{4}{3}\pi h^3 \rho_c$ and we approximate $p \approx p_c$ and $\rho \approx \rho_c$. (Because the pressure and density reach a maximum at the center, they do not change much from $r = 0$ to $r = h$, so these approximations are reasonably good if we take h small.)

Then we step outwards as we have done in Atmosphere. There are more accurate ways of starting out at $r = h$, but provided h is small enough our method is good enough. The interested reader is encouraged to try to invent better ways and to test them.

The program on the website incorporates the changes I have described, and with its comments it should be relatively straightforward to understand.

One sees from the figure that the model's pressure fits reasonably well overall, but that it is too low in the central region. A graph of density would show a similar trend. The temperature is not so good, being overestimated everywhere in our model by a factor of about two.

These inaccuracies are caused by two things. First, the composition of the Sun changes from inside to outside, because nuclear reactions are generating much heavier elements in the center; in the model we have assumed that the mean molecular weight μ in Equation 8.13 on page 92 is everywhere equal to its value at the surface. The surface value is 1.285, while the Standard Model takes $\mu = 1.997$ at the center. The second complication is that the effective value of the index n should be allowed to change with radius, partly because we have left out radiation pressure from the photon gas, and partly because the gas in the Sun outside $0.74R_\odot$ is in steady convection, bubbling outwards and then sinking downwards in a long, slow rolling motion.

Despite the differences, the agreement between our simple model and the very much more elaborate Standard Model is gratifying: considering that the Sun's central temperature is more than 1000 times larger than the surface temperature, our overestimate by a factor of two is a relatively small error. Using only the most el-

Investigation 8.6. Using the computer to model the Sun

In using the computer program Star described in Investigation 8.5, one has to face a big difference from the planetary atmospheres case: we have no direct observations of the interior of the Sun, so we do not know directly what values to adopt for the pressure and temperature there. These are needed to start the calculation off. All we know is what the real radius and mass of the Sun are, and these are the *results* of the computer program. It might seem that we are stuck with a trial-and-error approach to finding the structure of the Sun: try a set (p_c, T_c) and use the program to find what they predict about M and R . Then choose different starting values to see if the results are closer to or further from the true values M_\odot and R_\odot .

In fact, there is a more systematic way to guide one’s choices of new values for p_c and T_c . If we look at the equation of hydrostatic equilibrium for the Sun,

$$\Delta p = -\frac{Gm(r)\rho(r)}{r^2}h,$$

then we may ask what happens if we take h to be R itself: take one giant step from the center to the surface. Then Δp will be the difference between the surface pressure, which is zero, and the central pressure p_c : $\Delta p = -p_c$. If we take the mass term $m(r)$ to be the total mass M , the density $\rho(r)$ to be p_c , and r to be R , then although the equation is not accurate, it gives us a starting approximation for p_c :

$$p_c = \frac{GMp_c}{R}.$$

Next, we treat the mass equation, Equation 8.16, the same way. If we imagine that the central density is the density everywhere, then one jump from center to surface gives

$$M = \frac{4}{3}\pi R^3 p_c. \quad (8.18)$$

By solving this for p_c and substituting it into the equation for p_c we get

$$p_c = \frac{3GM^2}{4\pi R^4} \propto \frac{M^2}{R^4}. \quad (8.19)$$

I have dropped all the constants in the second form and simply written it as a proportionality, because with all of our approximations we cannot trust the actual value this equation will produce. But it does tell us something extremely important: given two stars with the same polytropic equation of state, we can expect that their central pressures will scale with mass and radius in approximately the way given by Equation 8.19.

The ideal gas law tells us how the central temperature behaves, and with the same approximations it gives

$$T_c \propto \frac{M}{R}. \quad (8.20)$$

This very simple equation helps us treat the central pressure in the same way as the central density.

Now we can see how we can correct erroneous values of the central pressure and temperature. Suppose we start with central values p_1

and T_1 , and suppose they give a model with mass M_1 and radius R_1 . We want to find starting values p_2 and T_2 that will give us the right values M_\odot and R_\odot . We write down the relevant proportionalities:

$$p_1 \propto \frac{M_1^2}{R_1^4}, \quad T_1 \propto \frac{M_1}{R_1}, \quad p_2 \propto \frac{M_\odot^2}{R_\odot^4}, \quad T_2 \propto \frac{M_\odot}{R_\odot}.$$

Assuming that the constants of proportionality are the same in each case (a big assumption: we will come back to this), we can divide the second set of equations by the first to obtain

$$p_2 = p_1 \left(\frac{M_\odot}{M_1} \right)^2 \left(\frac{R_1}{R_\odot} \right)^4, \quad T_2 = T_1 \left(\frac{M_\odot}{M_1} \right) \left(\frac{R_1}{R_\odot} \right). \quad (8.21)$$

This allows an *intelligent* correction of the first guesses for p_c and T_c , and can be used over and over again until the values of M_1 and R_1 converge to M_\odot and R_\odot , respectively. But will it work? After all, the approximations don’t seem very convincing: maybe the “constants” of proportionality in Equations 8.19 and 8.20 will depend on the structure of the star in some way that will make them change with changes in the central values of p and T , thus invalidating Equation 8.21.

The general answer is that if the first values p_1 and T_1 are pretty close to the right ones, then the constants of proportionality can’t change much, and the corrected values of p_c and T_c will be better than the old ones. The procedure will close in on the right model if we repeat the corrections often enough, provided we start close enough to the right answer in the first place.

The proof of a pudding is in the eating. Try the method on the computer model. (Don’t use the values of p_c and T_c supplied in the computer program, since they are the “right” ones.) You will be pleasantly surprised: you will find you need only *one* correction to get very close to the right mass and radius, no matter how far from the correct values you start! The method actually works better than we should expect. The reason is an “accident” that we have not made use of: for polytropes, the proportionalities in Equations 8.19 and 8.20 are in fact strict proportionalities, provided one keeps the polytropic index fixed.

The last open question is, what is the right index for the polytropic equation of state? We cannot answer this here: the Sun is not really a polytrope anyway. All we can do is find a polytropic index that comes close to the structure of the Standard Model. After some experimentation, I have settled on the value of 2.8 for the variable called index. For a model with the solar mass and radius, this gives a pressure curve that is quite close to that of the Standard Model, and a temperature curve that is usually within a factor of two of the standard one. This is about the best one can do with a single polytropic equation of state valid everywhere. The graphs in Figure 8.3 on page 92 show that our model is not a bad representation of the Sun, but it is clear that we would have to put in more physics to get all the detail right.

ementary techniques, and only observed data at the surface of the Sun (its radius, mass, and composition), we have learned quite a bit about the unseen regions of the Sun.

How photons randomly ‘walk’ through the Sun

Let us now put together two facts of common experience to learn a little about what happens to photons inside the Sun. The first fact is that the light that comes from the Sun’s surface is visible light, with some ultraviolet. It can burn our skin, but it isn’t strong enough to get inside our bodies, like X-rays do. The second fact is that radioactivity commonly produces *gamma-rays*, which are light waves that have more than enough energy to damage the insides of our bodies. Now, if the Sun is powered by nuclear reactions, then it seems sensible to assume that the nuclear reactions in the Sun’s center also produce gamma-rays. Why then do we get visible

In this section: remarkably, it is not easy for a photon generated deep inside the Sun to get out: it takes millions of years. We show that this is due to random scattering from gas particles. We construct a computer program to describe what mathematicians call a random walk.

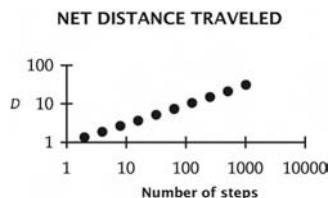


Figure 8.4. The net distance traveled in a random walk, as a function of the number of steps in the walk. The horizontal scale shows the number of steps, and the vertical scale is the average distance from the origin that the walk finishes, in units of the average length of each step of the walk. This is the output of the computer program Random from the website.

►Compton scattering is named after its discoverer, the American physicist Arthur Holly Compton (1892–1962). The phenomenon of scattering, in which a photon can lose or gain energy, is a further demonstration that light behaves like particles as well as like a wave.

light from the surface, and not gamma-rays?

The answer has to do with the photons' energy: the energetic gamma-rays produced inside are somehow losing energy before they reach the surface. The only way this can happen is through scattering: photons scattering from electrons and getting into temperature equilibrium with them. One of the biggest over-simplifications of our computer model is that we have neglected the transport of photon energy outwards through the Sun. In fact, getting the photon energy to the surface to be radiated away is a surprisingly long and tortuous process.

We will come back to the nuclear reactions that generate the Sun's energy in Chapter 11, where we discuss them in the context of all stars. Here we simply assume that energy in the form of photons is being released in the center: what happens then?

The radiation at the center of the Sun is very energetic, having been produced as gamma-rays. Each photon has millions of electron volts of energy when it is produced. But the gas in the Sun is really a plasma of individual charged particles, and photons scatter off charged particles very easily. This scattering is called **Compton scattering**, and the resulting exchanges of energy between electrons and photons lead to thermal equilibrium between matter and radiation: both have the same temperature.

If we want to decide how important scattering is inside the Sun, the key question is, how far can a photon travel between one scattering and another? The answer is surprisingly easy to work out from some simple basic numbers, and we do this in Investigation 8.7. We see there that the typical distance that a photon can travel in the Sun before it Compton scatters off another particle is no more than about 3.6 cm, which is a fraction 5×10^{-11} of the radius of the Sun!

If the photon were to travel on a straight line from the center to the surface, it would scatter 2×10^{10} times before emerging. This is certainly a lot of scatterings, so it is not surprising that the photon loses energy as it goes along. But in fact it cannot move on a straight line, since every scattering changes its direction of travel. The photon executes what mathematicians call a *random walk*, moving in random directions with steps of average length 3.6 cm. We show, using the simple computer program Random, described in Investigation 8.7, how a photon makes gradual progress outwards in this random, aimless way.

In fact the photon must scatter about 2×10^{10} squared times – 4×10^{20} times! – before it reaches the surface of the Sun. In doing so it will travel a total distance equal to 2×10^{10} times the radius of the Sun in order to get out, which at the speed of light takes more than a thousand years! Photons don't stream outwards; they diffuse very gradually.

Of course, it over-simplifies matters to imagine that a photon retains its identity all the way along this walk. In fact, photons are often absorbed by ions, and new ones are sometimes generated when charged particles collide. But the calculation still tells us how long it takes the energy carried by photons to diffuse outwards.

One effect of all this scattering is that, wherever the photon finds itself, it will be part of a photon gas that is in temperature equilibrium with the gas of the Sun. All the initial energy of the photon is lost quickly, and it adopts the energy of the particles (the free electrons and ions) that it is scattering from. Only at the very surface of the Sun does the probability that the photon will escape without a further scattering become large. This surface of last scattering is called the *photosphere* of the Sun, and the photons that come to us from it have the energy of the gas there, not the energy they started with at the center.

Investigation 8.7. The aimless walk of a photon through the Sun

The Sun is a dense cloud of electrons and ions, so it is not an easy place to be a photon. Photons scatter from charged particles, and ignore electrically neutral particles. Photons will also scatter from neutral *atoms*, because they actually encounter the electrons orbiting around the nuclei of the atoms, and they scatter from these.

It is not hard to estimate how far a photon can go before it scatters. To do so, we need two numbers: how many scatterers there are in a given volume of the Sun, and how “big” a scatterer is. The photon’s problem is a bit like that of the ball in a pinball machine: if the scatterers are big enough, and if there are enough of them, then the photon can’t go far without running into one.

We shall get a minimum estimate of how much scattering takes place by assuming that the photons scatter only from electrons and protons. In fact, in the Sun (as in most stars), ions of other elements contribute a very large amount to the scattering. Scientists use the word *opacity* to describe the amount of scattering, and ions of elements heavier than hydrogen and helium provide most of the opacity in the Sun. So our calculation here sets a lower limit on the opacity.

Assuming our scatterers are just electrons, what is their “size”? This size really refers to a kind of sphere of influence: how close a photon can get to the electron before it has to scatter. In quantum theory, the electron is not a solid particle of fixed size, but it does have a well-defined range of influence on photons, which is given roughly by what physicists call the “classical electron radius”. Its value is about $r_e = 2.8 \times 10^{-15}$ m. (See the Appendix for a more accurate value.)

This means that an electron presents an area to the photon equal to the cross-sectional area of a sphere of the same radius, which is $2\pi r_e^2$. (This underestimates the actual effective cross-sectional area for scattering by about 2/3, but this is close enough for our calculation.) If the photon comes within this “target” area around the position of the electron, it will scatter strongly. If it passes further away, it may still scatter more weakly, but we will not make a huge mistake if we just treat the electron as a solid target of radius r_e .

The number of electrons per unit volume in the Sun is easy to calculate: the Sun is mainly hydrogen, which has just one electron and one proton per atom, and almost all of these atoms are actually ionized: the electrons and protons are separated. Therefore, the number of electrons equals the number of protons. Essentially all the mass of the Sun is in its protons, since the mass of an electron is only about 1/2000th of the mass of a proton. The number of protons in the Sun is then roughly the mass of the Sun divided by the mass of a proton: $N_p = M_\odot / m_p$. Looking up these numbers in the Appendix, we find $N_p = 1.2 \times 10^{57}$. This is also then the number of electrons in the Sun. The average number per unit volume is this divided by the volume of the Sun. The Sun’s radius is $R_\odot = 7 \times 10^8$ m, so its volume is $V_\odot = 4\pi R_\odot^3 / 3 = 1.4 \times 10^{27}$ m³. The average number of electrons per unit volume, n_e , is the ratio: $n_e = N_e / V_\odot = 8.4 \times 10^{29}$ electrons per cubic meter.

We now calculate the average distance a photon can travel before it encounters an electron. Imagine the electrons as being solid balls of radius r_e , distributed randomly around the present location of our photon. If the photon moves in some directions, it will immediately run into an electron. In other directions, it will miss the nearby ones and travel a larger distance before hitting one. Imagine drawing a sphere around the photon’s present location. If this sphere is sufficiently small, the photon will have a pretty good chance of reaching it without hitting an electron. If the sphere is large, the photon will almost certainly hit an electron before it reaches the sphere. The sphere which the photon has roughly a 50-50 chance of reaching before it encounters an electron must be the one which contains just enough electrons that their cross-sectional areas equal the area of the sphere. We could just cover the inside of such a sphere if we arranged the electrons uniformly. In fact, they are arranged ran-

domly, overlapping in places and leaving gaps elsewhere, so all we can say here is that this sphere is about the right size for a ray randomly directed outwards from the center to have a good chance of hitting an electron no matter what direction it takes.

Now, protons also scatter photons, and since the proton charge is the same as the electron charge, except for sign, protons scatter photons just as well as electrons. It follows that the number of scatterers inside the sphere is actually twice the number of electrons. (Remember, we are ignoring complications due to ions, which in fact provide much more opacity than electrons and protons.)

Our argument tells us that the average distance the photon can go before scattering, ℓ , is roughly the radius of this sphere. The number of scatterers inside the sphere, E , satisfies $4\pi\ell^2 = 2\pi r_e^2 E$, or $E = 2(\ell/r_e)^2$. On the other hand, we know how to find E from the volume of the sphere and the number of electrons per unit volume: $E = 2n_e \times 4\pi\ell^3/3$. Setting these equal gives an expression that can be solved for ℓ :

$$\ell = \frac{3}{4\pi n_e r_e^2} = 3.6 \text{ cm.} \quad (8.22)$$

This is called the *mean free path* of a photon in the Sun. It is a fraction $\ell/R_\odot = 3.6 \times 10^{-2} / 7 \times 10^8 = 5 \times 10^{-11}$ of the radius of the Sun.

Now assume that the photon moves outwards by taking steps of this size, but in random directions: after each scattering its direction is different in a random way. We can find out the effect of this by writing a simple computer program to simulate such a random walk.

The program Random on the website uses the fact that computers are good at generating random numbers. (In fact, since nothing inside a computer is really random, computers use clever tricks to generate what are called pseudo-random numbers, which for most purposes can be used as if they were truly randomly chosen.) By choosing random steps in each of the three coordinate directions, we can simulate the aimless motion a photon in the Sun.

We want to know how far from the center a photon will get after a certain number of random steps. The program selects the number of steps and then performs a large number of trials for “walks” of this number of steps, calculating the average distance from the center at the end of each walk. It also calculates the average distance the photon goes in one step. In Figure 8.4 we plot the results. The axes of the graph are both logarithmic, which means that the linear distance increases uniformly for each factor of ten increase in the variable. This is ideal for showing relationships where one variable is a power of another. In this case, the graph shows that the average net distance of the walk, D , after N steps is $N^{1/2}$ times the average length of one step:

average finishing distance of a random walk of N steps

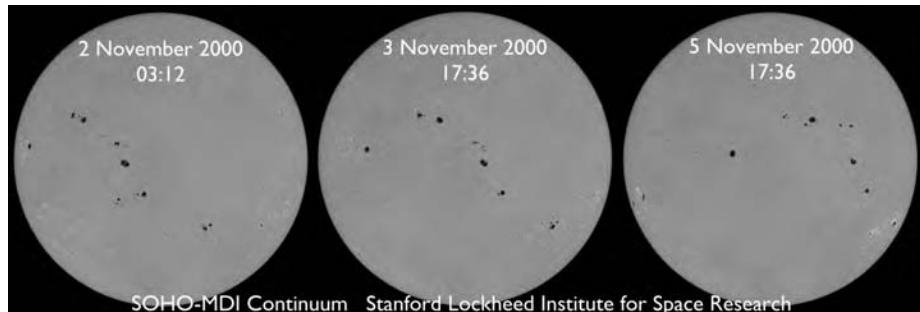
$$= \sqrt{N} \times \text{average length of one step.}$$

To see this, try a few cases. When the horizontal variable is 100 (a walk of 100 steps), a line through the points would give a vertical value of 10 (a net distance of 10 steps). When the horizontal variable is 1000, the vertical value is about 30, which is as close to the square-root of 1000 (31.6) as one can estimate from this graph.

So we have learned how the photon makes its way out: if it wants to reach the surface, which we have seen is equivalent to 2×10^{10} steps distant from the center, then it will have to make this number *squared* of actual steps to do so. This is 4×10^{20} steps of 3.6 cm each, a total walk of 1.4×10^{19} m, or 2×10^{10} times the radius of the Sun. At the speed of light, this takes 1500 years.

In reality, scattering from other ions makes the mean free path of a photon much smaller, so that the time it takes photons to emerge from the Sun is more like a million years!

Figure 8.5. The rotation of the Sun can clearly be seen from the motion of sunspots. Here, a sequence of images made by the MDI instrument on the SOHO satellite shows that the pattern of sunspots rotates nearly uniformly around the Sun. Adapted from images courtesy NASA/ESA.



The surface of the Sun is by no means as simple as our solar models would suggest, with the pressure and density going smoothly to zero. The outward streaming of radiation, the presence of magnetic fields, and the outward flow of pressure waves generated inside the Sun (see the section on solar seismology below) all conspire to produce a very complex region. The solar corona is a kind of atmosphere for the Sun. It extends far outside the Sun and, while being very rarefied, is also very hot. (Yet another place in the Solar System where the temperature begins to increase outwards!) Leaving the Sun is a constant stream of particles, called the *solar wind*. They flow outwards through the Solar System, disturbing the environments of all the planets. On the Earth, the very energetic particles produced by solar magnetic storms produce the aurora borealis and aurora australis phenomena.

Rotation keeps the Sun going around

In this section: we learn that the Sun rotates, which causes it to be slightly non-spherical. Sunspots give the evidence.

►Sunspots are places where the tangled magnetic field of the Sun pokes out of its surface. They come in pairs, like poles of a magnet.

Every 11 years the Sun's magnetic field reverses, with North becoming

South and vice versa. Sunspot numbers wax and wane on the same 11-year cycle.

►The other way of making an ellipsoid, with the long axis joining the poles, produces a football or egg shape, called a prolate ellipsoid.

Some galaxies are thought to be prolate, and prolate shapes can arise briefly when stars collapse, but the oblate form is much more common in the Universe.

In an earlier section I said that the Sun would collapse toward a single point if it were not for the pressure that holds it up against gravity. How, then, do the planets stay "up" in their orbits against the gravity of the Sun? After all, they are not affected by any significant pressure from the Sun. The answer is, of course, easy: they rotate about the Sun. Evidently, rotation can be an important source of support against gravity, so it is time we discussed it in the context of the Sun.

The evidence that the Sun does rotate is dramatic: **sunspots** migrate across the face of the Sun and often circle completely around it, returning for a second time before disappearing. This is illustrated in Figure 8.5. The motion of the spots is always the same, and the period of their rotation is always the same, regardless of whether the spots are large or small. Its cause is therefore not to be found in the spots themselves, but in the rotation of the Sun. Its rotation period is about 30 Earth days at the equator. The pole of the rotation axis is well aligned with the pole of the rotational motions of the planets, and the Sun's rotation is in the same sense as that of the planets.

Rotation should in principle change the shape of the Sun. At the solar equator, rotation will contribute a **centrifugal effect** that will bulge the equator outwards. At the pole, there is no such effect. We would therefore expect the Sun to have an elliptical shape, with the long axis in the equator. Such a three-dimensional elliptical shape – like a jelly doughnut – is called an oblate ellipsoid.

The Sun's rotation period is very long, however, in terms of the amount of rotation that would be needed to make a significant distortion. If there were a hypothetical planet orbiting the Sun immediately above its equator, it would have an orbital period of only 2.8 h. Put another way, if the Sun spun with a rotation period of 2.8 h, then it would begin throwing material off from its equator. Since its actual period is around 30 days, it is rotating very slowly.

How much distortion would we expect from this rotation? The shape of the Sun

can be measured by, say, the ratio of the minor to the major axis of its elliptical shape. This is a dimensionless number, so we might expect this to depend on the rotation rate of the Sun through another dimensionless ratio. The only one available is the ratio of the actual rotation rate to the maximum possible rotation rate. When the ratio is zero (no rotation) the distortion is also zero. So one might guess that the distortion would be proportional to the ratio raised to some exponent.

There is a simple argument that the exponent cannot be one: if it were, then it would change sign if the Sun rotated the other way, because this would give the Sun a negative period relative to the planets. But changing the sense of the Sun's rotation cannot change its shape, so the shape must depend on the second power of the ratio.

The next simplest guess is that the distortion will be proportional to the *square* of the ratio of these two rotation speeds, and more detailed calculations show that this is in fact correct. Since the ratio of speeds is of order 1/300, we should expect the expansion of the equator of the Sun to be of the order of 10^{-5} of its radius. This is exceedingly difficult to measure, since the edge of the Sun is not very well-defined: its brightness decreases gradually near its edge, not sharply. But recent observations confirm that the distortion is of this order.

Solar seismology: the ringing Sun

Observations of the Sun that we have encountered so far tell us about either the surface of the Sun – its brightness, size, composition, temperature, rotation, and surface magnetic field – or the very center where nuclear reactions are taking place. What about the vast, relatively inactive region in between? Until recently, we had no observational information about this region. The science of solar seismology, called **helioseismology**, is changing all that.

The Sun isn't simply the quiet ball of gas that our equilibrium model calculates. The energy generated in the center and the convection of that energy outwards produces, at some depths, a slow rolling of gas out and back again. This motion disturbs the surface of the Sun, producing small motions that can be detected by specially constructed solar telescopes. These observations have the potential to tell us as much about the interior of the Sun as studies of seismic waves have told us about the interior of the Earth.

The key to grasping the importance of these observations is to understand that the Sun has certain **characteristic frequencies** of vibration, just as does any other physical system, such as a violin string, a drum, an organ pipe, a bell, or a half-filled soft-drink bottle. It will repay us to think a little about the characteristic frequencies of such systems.

The set of frequencies of vibration of a violin string (the *acoustic spectrum* of the string) is relatively simple, consisting of the **fundamental frequency** (lowest frequency) and its **overtones**, which are just simple multiples of the fundamental frequency. Associated with each frequency is a pattern of vibration: at the fundamental frequency, the string vibrates as a whole. At the first overtone, which has twice the frequency, the string vibrates in two halves; if one half is mov-

▷This kind of argument, based on changing signs, may seem surprising when you first encounter it, but it can be a powerful guide to understanding the solutions to many kinds of problems.

In this section: the Sun has characteristic frequencies of vibration, just like any other body. The turbulent flow of energy out from the center excites these vibrations, and astrophysicists measure their frequencies. These are the best data we have about the nature of the interior of the Sun.

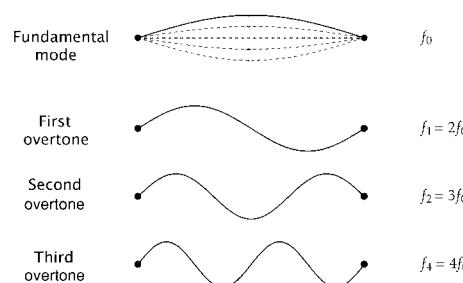


Figure 8.6. The first few characteristic frequencies of vibration of a string and the vibration patterns – normal modes – associated with them. The string is held fixed at its end points, as on a violin. When the string is set in motion with exactly the pattern shown, then it will move up and down with that same pattern, as shown for the fundamental mode. For the other modes, only the shape at one moment is shown. The fundamental frequency, f_0 , depends on the length, thickness, and tension of the string.



Figure 8.7. The classic example of the importance of normal modes in engineering is the Tacoma Narrows bridge disaster. The bridge was built across a waterway in the American state of Washington. Wind blowing across the bridge excited a normal mode whose pattern was a torsional oscillation of the roadway. The bridge became a tourist attraction on windy days, until a particularly strong wind drove the oscillations to a larger size than the structure could cope with. The roadway broke up on 7 December 1940. Reproduced with permission of the Smithsonian Institution.

ing upwards, the other is moving downwards, so that the midpoint of the string does not move at all. Such patterns are called the **normal modes** of the string.

When the string is bowed or plucked, the resulting motion is usually a combination of many normal modes, so that the sound produced includes several frequencies. For the ear, frequencies that are an octave apart (i.e. where one frequency is just twice the other) sound harmonious when heard together. Therefore, the string is ideal for making musical instruments: its fundamental and first overtone are separated by an octave. The fundamental and overtones of open and closed pipes have a similarly pleasant sound when heard together, so such pipes form the basis of organs and wind instruments.

Membranes, from which drums are made, do not have harmonious overtones. Therefore, drums are made so that either the normal modes damp out (decay) rapidly, producing a dull “thud” as in a typical bass drum, or the membrane is stretched over a “kettle” that resonates with and amplifies the fundamental frequency more than the higher overtones, as in orchestral tympani drums.

The analysis of normal modes is an important part of many other aspects of modern life. Engineers must routinely assess the frequencies and vibration patterns of all sorts of structures, including skyscrapers, road bridges, automobile chassis, aircraft bodies and components, and so on. All must be checked to see that the patterns of vibration are acceptable; that the characteristic frequencies are not likely to lead to the mode being amplified by external forces on the object (see Figure 8.7); and indeed that the structure is *stable*, which means that small disturbances of the structure will not spontaneously grow larger and larger.

The Sun vibrates in a way that is similar to other mechanical systems. We can deduce its typical vibration frequency from a simple argument. We saw in Equation 7.9 on page 79 that the square of the speed of sound in a body, including the Sun, is proportional to its pressure divided by its density. We also argued that this is similar to the square of the random speed of atoms in the Sun. But we know that this is also approximately kT/m , where m is the mass of the atoms (or ions). We also know that the average temperature of the Sun is not high enough to give the atoms the escape velocity, but it is not much lower either, so we can put this chain of argument together and roughly say that

$$v_{\text{sound}}^2 \propto v_{\text{escape}}^2 \propto GM_{\odot}/R_{\odot}.$$

Now, the frequency of vibration of the Sun must have to do with sound waves crossing the Sun back and forth in a regular pattern. The time it takes to cross is $R_{\odot}/v_{\text{sound}}$, and the frequency f is the reciprocal of this. This leads to the very important relation

$$f^2 \propto GM_{\odot}/R_{\odot}^3 \propto G\rho_{\odot}. \quad (8.23)$$

The fundamental frequency of vibration of the Sun is proportional to the square-root of its average density. When one puts numbers from the Appendix into this, one finds $f \approx 0.6 \text{ mHz}$.

Notice that the fundamental frequency of vibration is similar to the orbital frequency of a satellite at the surface of the Sun. The orbital speed is given by $v_{\text{orb}}^2 = GM_{\odot}/R_{\odot}$. The orbital frequency f_{orb} equals the circumference of the orbit divided by v_{orb} , which implies

$$f_{\text{orb}} \propto GM_{\odot}/R_{\odot}^3. \quad (8.24)$$

Thus, both the orbital frequency and the vibration frequency depend just on ρ . This is not a coincidence. Gravity is at work in both, fixing not only the orbital speed but also the pressure required to hold up the Sun, and therefore the sound speed.

Investigation 8.8. Making sure the Sun lasts a long time

We know from the age of the Earth that the Sun has been around a long time. It must therefore be stable, in other words resilient in its response to disturbances. Here we shall show that the great age of the Sun tells us that its polytropic index should be less than three.

The argument, like many that we have seen, is remarkably simple. We start with the rough solution of the structure equations for a star, Equation 8.19 on page 95, the most important part of which we reproduce here:

$$p_c \propto \frac{M^2}{R^4}.$$

We have already noted that this proportionality is strictly true if the star is a polytrope. Now consider a *sequence* of stellar models, all of which have the same composition and equation of state (so the constant of proportionality in this equation is the same) and the same mass. The members of the sequence will differ from one another in their central pressure and temperature, and their radii. If we fix the central pressure, we get a unique model, with a well-defined temperature and radius.

Since the masses of all the stars on our sequence are the same, we can write Equation 8.18 on page 95 as

$$M \propto p_c R^3 = \text{const.} \Rightarrow R \propto p_c^{-1/3}.$$

Replacing R in the equation for p_c by this, we find

$$p_c \propto p_c^{4/3}. \quad (8.25)$$

Along a sequence of stars of the same mass, the central pressure will be proportional to the 4/3 power of the central pressure.

What does this have to do with the ability of the star to resist a slight compression or expansion? The answer is that the balance between pressure and gravity that determines the structure of the

star is the same one that determines how the star will respond to a slight compression. The compression changes the gravitational field of the star by making the star more compact: gravity gets stronger. Compression of any gas also increases its pressure. If the compression produces more than enough extra pressure to resist the extra gravity, then the pressure will push the star out again. Such a star is said to be *stable*, and it will simply oscillate in and out, in one or more of its modes of oscillation. If on the other hand the star produces less extra pressure that is needed to resist the extra gravity, the star will continue to contract. Such a star is *unstable*, and even a very small compression will lead to its collapse.

Now consider a star just between these two cases: a star for which the pressure builds up exactly as much as is needed to compensate the extra gravity, and so the star remains just in equilibrium, neither bouncing back nor contracting further. Since the compression has not changed the mass of the star, compression must make it follow a sequence of equilibrium models of constant mass. We have seen above that along such a sequence the central density and central pressure are related by $p_c \propto p_c^{4/3}$. However, the equation of state of the star is, by hypothesis, a polytrope of the form $p \propto \rho^\gamma$. It follows that *if the equation of state has a polytropic exponent $\gamma = 4/3$, the star will remain in equilibrium when compressed*.

What about other stars? If the polytropic exponent γ exceeds 4/3, then the pressure increases faster for a given compression (a given change in ρ) than for the case of 4/3. Such a star will bounce back from compression, and so is stable. Conversely, if γ is less than 4/3, the star is unstable. If we re-express these results in terms of the astronomers' polytropic index n , as defined by Equation 8.12 on page 92, then the case of marginal stability ($\gamma = 4/3$) is $n = 3$. Models with $n < 3$ are stable, those with $n > 3$ unstable. Our solar model, with $n = 2.8$, is stable, as we expect from its long life.

Like musical instruments, the Sun will have an acoustic spectrum of characteristic frequencies, but these will in fact be much more complicated than those of any musical instrument. This is due to two factors: first, the Sun vibrates as a three-dimensional object, whereas most musical instruments use either one-dimensional vibrations (strings and air columns) or two-dimensional vibrations (drums, gongs, bells). Second, the Sun is held together by gravity.

In musical instruments and other everyday objects, there is usually a fundamental mode whose frequency is the lowest of all, and whose pattern of vibration involves the structure vibrating as a whole. Other modes have more complex patterns and higher frequencies.

The spherical oscillations of the Sun follow the same pattern: a lowest frequency and an ascending series above it. In Investigation 8.8 we discuss the forces that drive the mode with the lowest frequency. There we show that if the Sun is a polytrope with an exponent γ in the equation of state $p = K\rho^\gamma$ that is larger than 4/3, then it is stable: a spherical disturbance will make it oscillate rather than collapse or explode.

But when we look at modes that are not spherical, where, for example, the Sun is dimpling in somewhere and out somewhere else, the story is more complex. The fundamental mode still exists and has a pattern in which all layers of the Sun move together, but its frequency is actually in the middle of the spectrum. Above the fundamental is a whole series of modes (called pressure modes, or *p-modes*) of ascending frequency that resemble the pattern in terrestrial objects, but in addition there is a second series of modes *descending* in frequency, which are associated with buoyancy motions of different layers of gas in the Sun. These are called gravity modes, or *g-modes*. In both sequences, the frequencies are not "harmoniously" related; the exact values of the frequencies depend on the detailed structure of the Sun. In addition, the rotation of the Sun changes the frequencies of those modes

whose patterns of vibration involve waves rotating one way or another around the Sun.

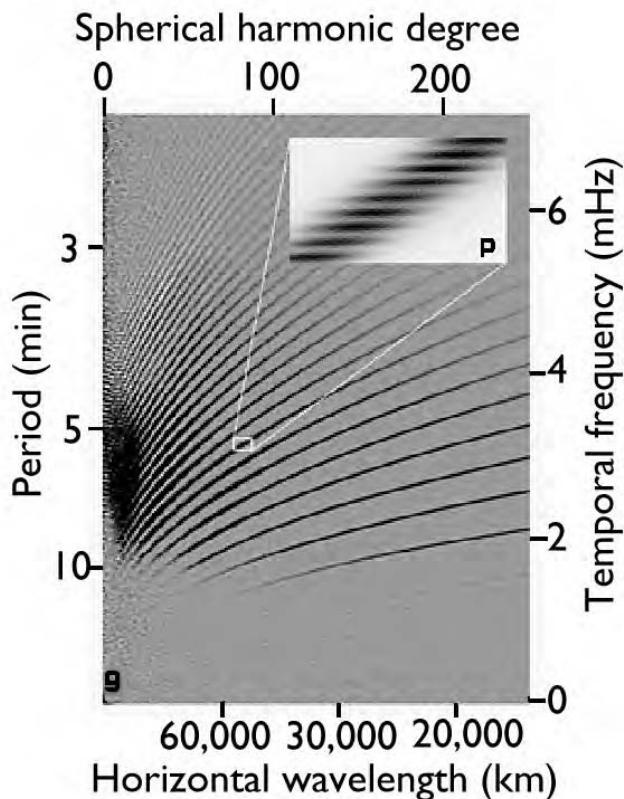


Figure 8.8. A chart of the frequencies of the p-modes of the Sun, as observed by the MDI instrument on the SOHO satellite. The horizontal axis is, as indicated, the horizontal wavelength of the waves, increasing to the left. The vertical axis is the frequency (increasing upwards) or the period (increasing downwards). The modes evidently form families. The g-modes, near the bottom of the chart, have not yet been observed.

Image courtesy NASA/ESA.

This complexity of the spectrum, and its sensitivity to the exact details of solar structure, explain why scientists are interested in measuring the frequencies of the Sun. By first constructing a numerical model of the Sun in great detail, then computing from it the frequencies of vibration that one would expect, and finally comparing those frequencies to the observed ones, one can test the accuracy of one's model. If the modes do not compare well, then the model can be changed until it reproduces the observed frequencies. The modes provide us with essentially the only way that we can "see" into the vast portion of the interior of the Sun in which nuclear reactions are not taking place. Figure 8.8 shows an example of the large amount of data that scientists have been able to gather.

The science of solar seismology is still young, but it has already provided corrections to the way the Standard Model of the Sun treats the flow of photons outwards and it has severely constrained solutions to the solar neutrino problem, which we will discuss in Chapter 11. Observations are continuing from satellites, like SOHO (see Figure 8.8), and from a number of ground-based observatories.

Reaching for the stars: the emptiness of outer space

With this chapter, we let gravity lead us out of the familiar territory of the Solar System and into the arena of the stars. This is a tremendous leap: the furthest planet, Pluto, is never more than 50 AU away from the Earth, while the nearest stars to the Sun – the α Centauri system – are 270 000 AU away! In between is almost nothing. Yet, just as gravity determines the structure of the Sun, so also it governs the stars.

Stars are the workplaces of the Universe. Stars made the rich variety of chemical elements of which we are made; they created the conditions from which our Solar System and life itself evolved; our local star – the Sun – sustains life and, as we shall see, will ultimately extinguish it from the Earth.

Leaping out of the Solar System

The huge variety of kinds of stars gives a clue to why they can do so many different things. There are stars that are 20 times larger than the whole Solar System, and others that are smaller than New York City. Big stars can blow up in huge **supernova** explosions; small ones can convert mass into energy more efficiently than a nuclear reactor. The material of which stars are made can take the form of a rarified gas thinner than the air at the top of Mt Everest. Or it can be so dense that ordinary atoms are squashed down into pure nuclei, so that a thimbleful of such material would contain more mass than a ball of solid steel 600 m across.

Stars affect each other in many ways. As they form together out of vast clouds of gas, most of them form pairs and triplets circling one another. The disturbances they produce in each other can lead to a range of fascinating phenomena, from **nova** outbursts to intense emissions of **X-rays**. Exploding stars can dump their debris into gas clouds that eventually form other stars. Some stars collide; a few even fall into black holes, with spectacular consequences. In all of these processes, gravity plays an organizing role; and all of them are important for understanding the Universe and indeed the origins of life on Earth.

Most important of all is that there are a lot of stars. The number we can see in the sky on the darkest of nights is a mere handful compared to the hundred billion stars that make up the collection that we call the Milky Way. There are so many stars that their mutual gravitational forces are strong enough to hold them together in a single spectacular **spiral galaxy** like that shown in Figure 9.1 on the following page. And there are perhaps a hundred billion such galaxies in the part of the Universe that we can study with our telescopes. That adds up to a lot of stars!

It is impossible to understand stars and the ways they affect one another unless one first comes to grips with the enormous distances between them. Ancient astronomers had some idea of the size of the Earth and the distance to the Moon. But all they knew about stars was that they were far away, very far away, too far to measure. Actually measuring the distance to the stars was one of the greatest steps in the development of modern astronomy. In this chapter, we shall concentrate on

In this chapter: how astronomers measure the brightness and distances of stars.

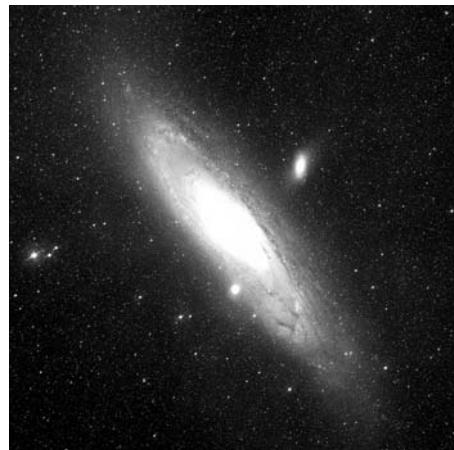
In this section: the huge number and variety of stars.

▷The biggest stars are called giants, and the smallest are neutron stars.

▷The image under the text on this page is a photograph of the sky showing the constellation of Orion, one of the easiest to recognize. In the sword, the group of three objects arranged in a roughly vertical line, the central fuzzy one is the Orion Nebula, which is a nursery where new stars are being formed. (See Chapter 12.) (Image copyright Till Credner, AlltheSky.com, used with permission.)

▷As a rule of thumb, a typical galaxy has 10^{11} – 10^{12} stars.

Figure 9.1. The spiral galaxy called the great galaxy in Andromeda, the Andromeda galaxy or M31, which is similar to our own Milky Way in size and shape. Like our own, it has several small satellite galaxies. If this were a photo of the Milky Way, the Sun would be about three-quarters of the way out in the disk. The view from that location would reveal an arc of stars circling the sky, with fewer stars in other directions. This is just what we see in our sky, and we call it the Milky Way. The Andromeda spiral is the nearest large galaxy to our own, and we are bound together gravitationally. In fact, we are approaching one another and will collide within a few billion years! Use of this image is courtesy of the Palomar Observatory and Digitized Sky Survey created by the Space Telescope Science Institute, operated by AURA, Inc. for NASA and is reproduced here with permission from AURA/STScI.



how this is done, and on what we can learn about stars as a direct result of measuring their distances. In the next chapter, we shall look at what actually goes on inside stars and how they are born and die.

How far away are the stars?

In this section: how astronomers know the distances to stars, and what they are. The parallax method is the most direct, but is only the first step on a complex distance ladder.

▷ Compare this use of parallax with that described in Figure 4.2 on page 26.

▷ An arcsecond is $1/3600^{\text{th}}$ of a degree, or 4.85×10^{-6} rad.

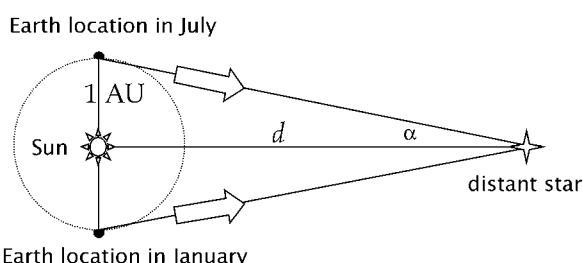
Astronomers measure the distance to the nearest stars by the same method that the ancient Greeks used to measure the distance to the Moon: triangulation. Where the Greeks took sightings on the Moon from different places on the Earth, modern astronomers take sightings on stars from different places on the Earth's orbit around the Sun. The change in the apparent position of a star when it is viewed from different places is called its parallax, and it is described in Figure 9.2.

Astronomers can measure parallax angles at least as small as one-tenth of an arcsecond. A star with a parallax of 1 arcsecond would be 206 000 AU distant. (This can be obtained from the formula in the caption of Figure 9.2 by setting α to 4.85×10^{-6} .) In practice, the nearest stars have parallaxes a bit smaller than this, but nevertheless this distance is fairly typical of the distances between individual stars. For this reason it has become a fundamental unit of length, and astronomers have given it a name: the **parsec**, the distance that gives a parallax of one arc second. It is abbreviated as pc, and its value is

$$1 \text{ pc} = 2.06 \times 10^5 \text{ AU} = 3.08 \times 10^{16} \text{ m} = 1.91 \times 10^{13} \text{ mi} = 3.25 \text{ light years.}$$

Astronomers express all cosmic distances in parsecs or in larger units derived from the parsec, such as the **kiloparsec** (abbreviated kpc) – about one-tenth the size of a typical galaxy – or the **megaparsec** (abbreviated Mpc) – typical of the distances between galaxies. We shall use these units too, since using meters or miles or other

Figure 9.2. The distance to a star can be determined from the change in the direction to the star as the Earth moves from one side of the Sun to the other. The different directions are indicated by the arrows from the Earth to the star in January, and then six months later in July. The parallax is defined to be the angle α indicated in the diagram. If α is measured in radians, then if it is small enough (always the case in practice) it is related to the distance d to the star to an excellent approximation by the equation $\alpha = (1 \text{ AU})/d$.



human-scale measures would only confuse us with the large powers of ten we would always have to use.

These distances are truly vast. The mass of the Solar System is almost all concentrated in the Sun, a ball of radius 7×10^8 m. The separations between stars are typically 10^8 times larger than this. The intervening space is largely empty: if we draw a box around the Sun whose sides are halfway to the nearest stars, then the Sun occupies only about one part in 10^{24} – one million-million-millionth – of the volume of this cube! In this vast space there is some diffuse gas (adding up to perhaps 10% of the mass of the Sun), but there may be as much as ten times the mass of the Sun in **dark matter**. This unseen substance is one of the great puzzles of modern astronomy. In Chapter 14 we will deduce its existence from its gravitational effects, but astronomers so far have not directly observed it. Yet even this dark matter is spread out over such a large volume that its density is unimaginably small.

In trying to measure stellar distances by parallax, astronomers have a problem: the twinkling of the stars. The Earth's atmosphere is turbulent, and light from a star has to pass through it before it reaches our telescopes. This turbulence causes the image of the star to jump around randomly. To the eye this causes the familiar and rather pleasant effect of twinkling. But to the astronomer, this is a nuisance, because it limits the accuracy of any single measurement of the position of a star from the ground to typically about one arcsecond.

The way to get around this problem is to make many measurements of the position of a star. By patiently performing hundreds of such measurements on the same star, astronomers can estimate the average position of the star and remove much of the confusion caused by twinkling. In this way, astronomers using telescopes on the ground have measured parallaxes smaller than 0.1 arcseconds, and so measured the distance to stars more than 10 pc away.

The obvious way to remove the twinkling problem completely is to make these measurements from a telescope in orbit about the Earth, above the disturbing effects of the atmosphere. A specially designed satellite called Hipparcos has done just that. Built and launched by the European Space Agency (ESA), it has measured parallaxes of thousands of stars to an accuracy of about 0.01 arcseconds or better. These data have greatly improved astronomers' understanding of many aspects of stars and their evolution. We will see an example in Figure 12.4 on page 140.

For stars that are more distant than we can reach even with Hipparcos, our estimates of distance are decidedly less accurate. There are many complications, but almost all methods use the brightness of a star. We can estimate the distance to a star by comparing how bright it appears with how bright we think it really is. The method is described in the next section.

How bright are stars?

The everyday term *brightness* has two different uses, so physicists use different words for them. The first sense of brightness is the total energy given off in a unit time. For example, a light bulb may be rated at 100 W. This means it gives off an energy of 100 J in each second. (For an ordinary tungsten incandescent bulb, most of that energy is in heat, and only a small part comes out as light; but all the energy comes out in one way or another.) Physicists call this the **luminosity** of the object, its total emission of energy in a unit time.

The other meaning is really the *apparent brightness*, which describes how a given star looks dimmer and dimmer as it gets further away. This happens because the total energy being given out in radiation is spread out over a larger and larger area as it moves out from the star. Any given observer gathers light from a fixed

►Astronomers call this twinkling effect "seeing", and they build telescopes in places where the seeing is very good – the twinkling is small.

In this section: we learn about the difference between the apparent brightness of a star and its absolute brightness, also called respectively the flux and luminosity of the star.

area, perhaps the area of the pupil of the eye, or the aperture of the telescope. The further away the star is, the smaller will be the fraction of the star's energy that will fall on a given detecting area, so the dimmer the star will seem to be.

Since the light from the star is spread over the area of a sphere surrounding the star at any distance, and since the area of the sphere is proportional to the square of the distance r to the star, the energy falling on a given detecting area decreases as $1/r^2$. Because less energy falls on the area as it moves further and further away, the star becomes dimmer and dimmer. If one imagines using a detector of a given unit area (say 1 m^2), then one can measure the energy falling on that unit area in a unit time. This is what the physicist calls **flux**: energy per unit time per unit area. So the apparent brightness of a star is called its flux, measured in joules per second per square meter, or W m^{-2} . For a "point" source of light, like a star, the apparent brightness is proportional to $1/r^2$.

Let us use these ideas to discover how luminous the nearest star to the Sun is: α Centauri. Its flux can be measured and turns out to be $2.6 \times 10^{-8} \text{ W m}^{-2}$. We saw earlier that this star is at a distance of $2.7 \times 10^5 \text{ AU}$, or $4.1 \times 10^{16} \text{ m}$. If we multiply the flux by the area of the sphere over which the star's light is being spread at this distance ($4\pi r^2 = 2.1 \times 10^{34} \text{ m}^2$), we get the star's luminosity: $5.5 \times 10^{26} \text{ W}$.

What happens if we do the same calculation for the Sun, using its measured energy flux (usually called the **solar constant**), 1355 W m^{-2} , and its distance, 1 AU or $1.5 \times 10^{11} \text{ m}$? We obtain the solar luminosity,

$$\text{solar luminosity} = 1L_\odot = 3.83 \times 10^{26} \text{ W}. \quad (9.1)$$

The luminosity of α Cen turned out to be 1.4 times the luminosity of our Sun. This is a reassuring result: we are not finding anything wildly different for our nearest star.

Astronomers' units for brightness

In this section: astronomers measure brightness in magnitudes, a logarithmic scale that goes backwards, so that the brightest stars have the lowest numbers.

Astronomers have evolved their own way of describing the luminosity and apparent brightness of stars. In astronomy books one does not find the conventional units of watts (W) or watts per square meter (W m^{-2}). Instead, one finds that the apparent luminosity of a star is described as its **apparent magnitude**, or simply its **magnitude**, and is called m . One also finds the total luminosity described as the **absolute magnitude** M .

The details of the definitions of these magnitudes are given in two analyses, first Investigation 9.1, then Investigation 9.2 on page 108, but there are two features that we should note here. First, magnitudes run the "wrong" way: stars with larger magnitudes are *dimmer* than stars of smaller magnitudes! And second, magnitudes run on what we call a **logarithmic scale**, which means that if the magnitude of one star is larger by one than the magnitude of another star, then the first star is dimmer than the second by a fixed *factor*, in this case about 2.512. A difference of two magnitudes implies a brightness ratio of $(2.512)^2$, or 6.31. The magnitude scale is arranged so that a difference of 2.5 magnitudes corresponds to a ratio of exactly 10 in brightness. As is described in Investigation 9.1, these peculiar brightness scales came about through a combination of historical practice and adaptations to the physical properties of the human eye.

On these scales, the star α Cen has an absolute magnitude M of 4.3 and an apparent magnitude m of -0.08. The Sun has an absolute magnitude M of 4.7 and an apparent magnitude m of -27. These numbers for absolute magnitude reflect the fact that the Sun is less luminous than α Cen by a factor of 1.4. The apparent magnitude numbers are dominated by the fact that the Sun is so much closer to the Earth than is α Cen, so it has a much greater apparent brightness.

Investigation 9.1. How ancient astronomers constructed their magnitude scale

In the days before there were any systematic units for physical measurements of energy or brightness, ancient astronomers needed to classify stars according to their relative brightness. It probably seemed natural to them to use a scale a bit like one they might have used for important people: the brightest stars were stars of the *first magnitude*, somewhat dimmer stars were of the second magnitude, and so on. Their brightness scale thus was the reverse of what would seem natural to us today: the larger the magnitude, the dimmer (less important) the star.

Ancient astronomers took the steps on the magnitude scale to correspond to levels of brightness that the eye could clearly distinguish. Moreover, in trying to keep the steps from one magnitude to another uniform, they devised a scale where given magnitude changes represent given *ratios* between brightnesses. Basically, this is because the eye and the brain are much better at saying "star X is twice as bright as star Y" than at saying "star X is brighter than star Y by the brightness of star Z". Here is how these two statements would get translated into measurement scales.

"*Star X is brighter than star Y by the brightness of star Z.*" Let us consider the second approach first. It is the one that the eye is not able to do well and hence does not correspond to the scale adopted by ancient astronomers, so we will be able to discard it. But it is important to understand it, since it would have led astronomers to the sort of scale that a modern physicist would try to devise.

Suppose the eye could in fact sense fixed brightness differences. Then ancient astronomers would have constructed their scale by deciding upon some chosen star as their brightness standard. Let us call this star Z, and denote its brightness by B_Z . (Since ancient astronomers would have had no way of relating this brightness to other brightnesses on the Earth, they might just have called this brightness 1.) They would then have found another star, say Y, that was brighter than Z by the brightness of Z itself, and they would have assigned this a brightness of $2B_Z$. They would have looked for brighter stars,

and found one, say X, that was brighter than Y by the brightness of Z: this would make $B_X = 3B_Z$. This is what we call a *linear scale*: the quantity used to describe the brightness changes from one star to the next in a way that is proportional to the change in the brightness itself.

Unfortunately, physical properties of the eye prevented ancient astronomers from using this sort of scale. If star X is 100 times the brightness of star Z, while star Y is 99 times as bright as Z, then the eye can see all three stars at once but it cannot even tell that X is brighter than Y, let alone that the difference is just the brightness of Z. So it is impossible to construct a linear scale for stellar brightness using the eye as a measuring instrument. Let us therefore consider the alternative.

"*Star X is twice as bright as star Y.*" This is the sort of statement that the eye *can* make fairly accurately. It can tell that, say, Y is twice as bright as Z, and that X is twice as bright as Y. It does not need to measure brightness differences to do this, only brightness *ratios*. Actually, what the eye is good at is telling that the ratio of brightness between Z and Y is the same as between Y and X: it is not so good at telling whether this ratio is exactly 2 or maybe 2.5 or 1.6 or something in between.

Ancient astronomers used this property to devise their magnitude scale. They took a given star, say X, as their standard and called it a star of the first magnitude (in modern language, $m = 1$). Then they found one, called Y, that seemed roughly half as bright as X and called it a star of the second magnitude ($m = 2$). They further found another star, Z, that had the same ratio to Y as Y had to X, and called that a star of the third magnitude ($m = 3$).

We call this a logarithmic scale because, as the table below shows, changes in our brightness scale (the magnitude) are proportional to changes in the *logarithm* of the brightness itself. Since the magnitude decreases as the brightness increases, the scale runs in reverse.

The ancient astronomers' magnitude scale

Hypothetical star	Brightness	Magnitude	Logarithm of brightness
Z	B_Z	3	$\log(B_Z)$
Y	$2B_Z$	2	$\log 2 + \log(B_Z)$
X	$4B_Z$	1	$2 \log 2 + \log(B_Z)$

Standard candles: using brightness to measure distance

Now, α Cen has other properties that astronomers can measure. For example, they can obtain its spectrum by dispersing its light through a prism or a diffraction grating. Although the general shape of the spectrum will depend mainly on the star's temperature (we will look at how this happens below), the details of the spectrum are a very sensitive "signature" of the star. Suppose an astronomer measures the spectrum of another star and finds that it is almost identical to that of α Cen. Experience has shown astronomers that the two stars will also be very similar in their other properties, such as their mass, radius, and total luminosity. If it turns out that the second star has a flux (apparent luminosity) that is only one quarter that of α Cen, then it is likely that this is because the star is twice as far away, so that its light is spread over a sphere whose area is four times as large as that over which α Cen's light is spread when it reaches us.

This is how astronomers estimate distances to stars further away than a few tens or hundreds of parsecs. It is almost the only way that distances can be measured until we reach the enormous distances that separate the giant clusters of galaxies in the Universe, where we can use the expansion of the Universe itself to measure distances. Obviously, the accuracy of the apparent-brightness method of estimating distances depends on how well we can estimate the total luminosity of the distant

In this section: astronomers estimate the distances to most objects from their apparent brightness. This works well if the intrinsic luminosity of the object is known. The search for such "standard candles" is one of the most fundamental activities in astronomy.

Investigation 9.2. How modern astronomers construct their magnitude scale

The modern definition of *apparent magnitude* is taken from the ancient one with only minor changes. The most significant change is that the ancient astronomers' estimate of a factor of 2 decrease in brightness for one step in magnitude was a bit low, and one can make a reasonable fit to the ancient magnitudes by taking the ratio to be nearer 2.5. In order to make things simple, astronomers define a ratio of brightness of 10 to be a change of magnitude of exactly 2.5.

In equations, this is fairly straightforward. Take two stars of brightness B_Z and $B_Y = 10B_Z$. The logarithms of their brightnesses are $\log(B_Z)$ and $1 + \log(B_Z)$. If the magnitudes are to decrease by 2.5 then we have

$$m_Y - m_Z = -2.5[\log(B_Y) - \log(B_Z)].$$

If we combine the logarithms into a single term, then we get

$$m_Y - m_Z = -2.5 \log\left(\frac{B_Y}{B_Z}\right). \quad (9.2)$$

This shows clearly that magnitude differences depend on the *ratios* of brightnesses. If two stars have a magnitude difference of 1, then Equation 9.2 shows that the logarithm of their brightness ratio will be $1/2.5 = 0.4$ and so their brightness ratio will be $10^{0.4} = 2.512$. This is close enough to the ratio of 2.5 mentioned at the beginning of this section.

The apparent magnitude scale is fixed if we adopt one star as a standard. Here we find the second difference between modern and ancient astronomers. Because the three brightest stars are in fact Southern Hemisphere stars, while the magnitude system was invented by Northern Hemisphere astronomers, the modern scale has to assign some stars to negative apparent magnitudes. The modern scale is chosen so that our old friend αCen is 7.6% brighter than a standard zero-magnitude star. This gives us an alternative to Equation 9.2,

$$\begin{aligned} m_{\text{star}} &= -2.5 \log\left(\frac{\text{Flux from star}}{\text{Flux from } \alpha\text{Cen}/1.076}\right) \\ &= -2.5 \log\left(\frac{\text{Flux from star}}{2.4 \times 10^{-8} \text{ W m}^{-2}}\right). \end{aligned} \quad (9.3)$$

Exercise 9.2.1: Magnitude of the Sun

Use the solar constant, given just before Equation 9.1 on page 106, to compute the apparent magnitude m of the Sun, using Equation 9.2. Use Equation 9.4 to calculate the absolute magnitude M of the Sun.

Exercise 9.2.2: Stellar magnitudes

A particular star is known to be ten times further away than αCen and five times more luminous. Compute its apparent and absolute magnitudes.

(See Table 10.1 on page 110 for a list of magnitudes of the brightest stars.)

Modern astronomers also know that stars are at different distances from the Earth, so that in order to understand them physically we need to measure their intrinsic brightness, or luminosity, and not just their apparent brightness. Astronomers have adopted a scale for the absolute magnitude M based on the apparent magnitude scale. The *absolute magnitude* of a star is numerically equal to the apparent magnitude it would have if it were 10 pc from the Earth.

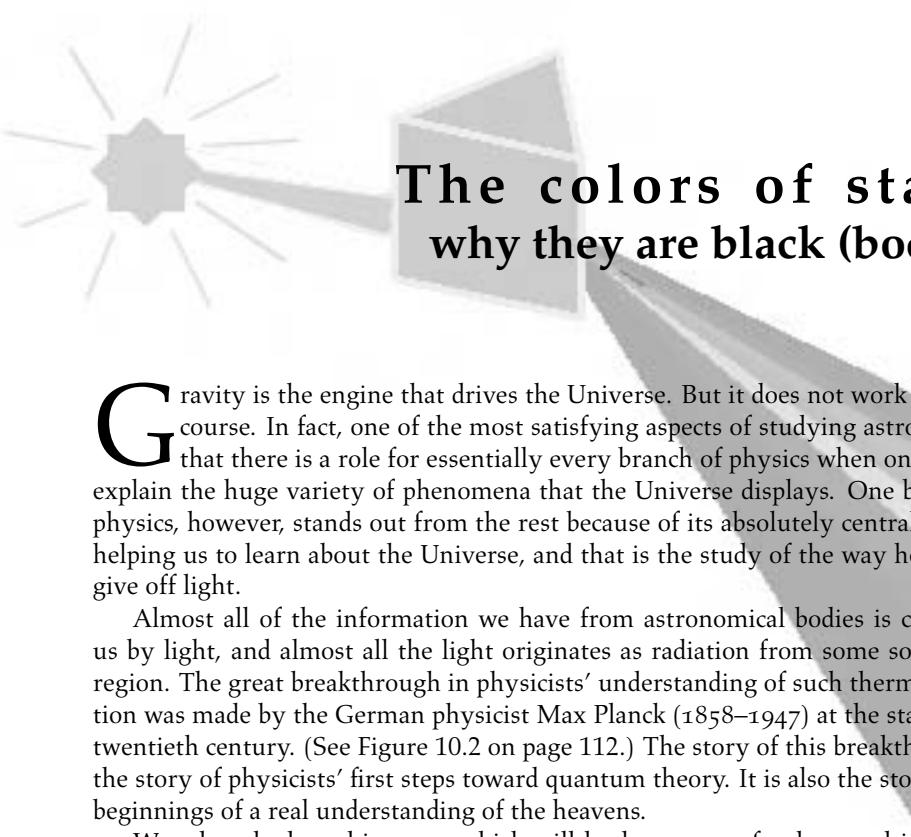
Since αCen is 1.33 pc away, we would have to place it 7.5 times further away to make its apparent magnitude equal its absolute magnitude. This would reduce its flux by $(7.5)^2 = 56.25$ and increase its apparent magnitude by $2.5 \log(56.25) = 4.38$. So the absolute magnitude of αCen is $-0.08 + 4.38 = 4.3$. A star with absolute magnitude 0 would have a luminosity larger than that of αCen by a factor of $10^{(4.3/2.5)} = 52$. This would give a luminosity $2.9 \times 10^{28} \text{ W}$, so that the absolute magnitude can also be written as

$$\begin{aligned} M &= -2.5 \log\left(\frac{\text{Luminosity of star}}{2.9 \times 10^{28} \text{ W}}\right) \\ &= -2.5 \log\left(\frac{\text{Luminosity of star}}{75 L_\odot}\right). \end{aligned} \quad (9.4)$$

Why have modern astronomers stuck to such an ancient and inconvenient system of magnitudes when most physicists have adopted more modern measuring scales for other quantities? The answer is at least partially that some astronomers are conservative and are reluctant to break the continuous tradition of astronomy that makes it the oldest of the mathematical sciences. But a much more important reason is simply that the logarithmic magnitude scale is useful. Stars and other objects come in a huge variety of luminosities and apparent brightnesses, and it is useful to have a scale where a change of magnitude of, say, 100 implies a brightness ratio of 10^{40} . This is big enough to span even the huge variations encountered in modern astronomy. Since it is useful to have a logarithmic scale for brightness, one might as well continue to use the ancient one!

star from other things we can measure about it. This depends on our finding similar stars whose luminosity is known, or can at least be calculated from some theoretical ideas about the object. Astronomers call such objects **standard candles**: objects whose intrinsic brightness is known.

In the past, astronomers have often changed their estimates of interstellar distances. However, in the last two decades of the twentieth century, some painstaking work with space observatories, coupled with a better understanding of important standard candles, has made astronomers' distance-scales much more accurate. Most astronomers now feel that their distance estimates, even over very large cosmological reaches, have errors no larger than 10%, and probably smaller.



The colors of stars: why they are black (bodies)

Gravity is the engine that drives the Universe. But it does not work alone, of course. In fact, one of the most satisfying aspects of studying astronomy is that there is a role for essentially every branch of physics when one tries to explain the huge variety of phenomena that the Universe displays. One branch of physics, however, stands out from the rest because of its absolutely central place in helping us to learn about the Universe, and that is the study of the way hot bodies give off light.

Almost all of the information we have from astronomical bodies is carried to us by light, and almost all the light originates as radiation from some sort of hot region. The great breakthrough in physicists' understanding of such thermal radiation was made by the German physicist Max Planck (1858–1947) at the start of the twentieth century. (See Figure 10.2 on page 112.) The story of this breakthrough is the story of physicists' first steps toward quantum theory. It is also the story of the beginnings of a real understanding of the heavens.

We take a look at this story, which will lead us to two fundamental ideas that together will unlock the secrets of a great deal of astronomy. These ideas are **black-body radiation** and the ionization of hydrogen. We will put them to work for us repeatedly in the next few chapters.

The colors of stars

We will start with the color of light. The different colors are, of course, just manifestations of the different wavelengths of light. The overall color of a body depends on the amounts of light of the various colors present in its emissions. Stars vary in color from red to blue, as you can easily see by using a pair of binoculars on a dark night.

Table 10.1 on the following page lists the magnitudes and distances of the five most prominent stars. Notice that the magnitude of α Cen is not quite what I quoted in the last chapter. This is not an error: it is because in this table I have used only the brightness of the stars in *visible* light. This is the so-called **visual magnitude** V of the star, and it is the most important measure of brightness for observations performed with the eye. Our previous discussion of brightness assumed we were dealing with all the radiated energy from the star, even radiation that comes out in the infrared or ultraviolet parts of the spectrum. This total emission is measured by the so-called **bolometric magnitude**, called M_b or m_b . The discrepancy in the magnitude of α Cen is due to the fact that some of its energy comes out in the ultraviolet and infrared: our earlier value was its bolometric magnitude, and Table 10.1 contains only its visual magnitude.

Stars emit more light at some wavelengths than at others. If a star is brighter at the blue end of the spectrum (short wavelengths) than at the yellow end (long wavelengths), then it will look blue, and if another star is brighter in the yellow than in the blue it will appear yellow.

In this chapter: the colors of stars give us insight not only into the stars themselves but into the branch of physics called quantum theory, founded by Planck and Einstein. The color of light tells us the temperature of its source because light comes in particles called photons. A star's color and brightness tells us its size and distance.

In this section: astronomers use standard filters to define the color of a star. Stars of different masses and compositions usually have different colors.

Table 10.1. Magnitudes of the five brightest stars. The first three are in Southern Hemisphere constellations. Most are nearby, but notice how far away, and how intrinsically bright, Canopus is!

Rank	Name	Constellation	m	M	d (pc)
1	Sirius	Canis Major	-1.46	1.4	2.7
2	Canopus	Carina	-0.72	-8.5	360
3	Rigel Kentaurus	Centaurus	-0.27	4.4	1.3
4	Arcturus	Boötes	-0.04	-0.2	11
5	Vega	Lyra	0.03	0.5	8.1

Astronomers have made this idea precise by defining a number they call the **color of a star**. This is based on measuring the magnitude of a star in different parts of the spectrum. If we filter the light through a blue filter before we measure the magnitude, we obtain the blue magnitude of the star, called B . If we filter through a filter in the central part of the spectrum, we get the visual magnitude V . These colors have become international standards, so that any astronomer wanting to measure B will use a filter that passes exactly the same range of wavelengths of light as any other astronomer would use. There are at least 11 such standard filters, ranging from the infrared nearly to the ultraviolet parts of the spectrum.

The nice thing about using a logarithmic scale for magnitudes is that the difference between any two magnitudes depends on the *ratio* of the brightnesses that the magnitudes measure – see Equation 9.2 on page 108. Therefore, if one takes the difference between any two filtered magnitudes, say $B - V$, one gets a number that measures the ratio of the blue brightness to the visual brightness. Now, since both brightnesses diminish with distance in the same way, their ratio is independent of how far away from the star we are.

Astronomers define the **color index** of a star to be the difference $B - V$. They measure the color index using only apparent magnitudes for V and B , not even needing to know how far away the star is.

Why stars are black bodies

In this section: black bodies absorb all light that falls on them, so stars are black bodies. Hot black bodies also radiate light, and they play a key role in quantum physics.

We saw in the last chapter that the color of the Sun tells us how hot it is, and that holds just as well for other stars. In fact, despite all the possible complications of stars – their size, their pulsations, their varying composition – there is a remarkably consistent relationship between the color and the temperature of a star. This is because stars are excellent examples of what physicists call **black bodies**! At first this seems like an outrageous abuse of common-sense language: how can a brilliant star be a black body?

The explanation is that the words “black” and “bright” actually refer to different physical processes. Physicists adopt the very reasonable definition that a body will be called a black body if it absorbs all light that falls on it. This applies not only to everyday blackness, such as black cats, black ink, or the black of the night, but also to stars. If we were to shine a light at the Sun, for example, the Sun would just swallow it up: no light would be reflected, and none would be transmitted through to the other side. So the Sun is, by this definition, black.

Our difficulty with this is that we are used to thinking that black objects are also *dark*: they do not shine. This is because in everyday circumstances, bodies either absorb or reflect light, and their color is determined by what wavelengths they reflect. If they are black (absorbing) they are also dark (sending no light back to us). However, bodies can also emit light all by themselves, and if they are sufficiently hot, we will see the emission. The burner of an electric stove starts out black, but when heated it glows red. For a physicist, it is still a black body: the black covering will still absorb any light that hits it. But in addition, it glows. Stars are the same.

Investigation 10.1. Black bodies

The spectrum of light emitted by a body is a measure of how much light comes out at different wavelengths. This is not quite so easy to define as it may at first seem, so here is how physicists do it.

- First of all, “how much light comes out” means the *rate* at which energy is being emitted in light: it is an energy per unit time.
- Next, the energy is radiated by the surface of the black body, and since each piece of the surface is independent of every other piece, the energy radiated will be proportional to the area. So physicists speak of the energy radiated per unit time and *per unit area*. Recall that in Chapter 9 we called this the *energy flux*.
- Of course, the energy comes out in various directions, so to avoid complications we will consider only the total energy that a piece of the surface radiates towards the outside of the body.
- Finally, the energy radiated may depend on the wavelength of the light. Since photons come out with a whole range of wavelengths, the chance of finding a photon with *exactly* some given value of the wavelength λ is essentially zero; it is more correct to speak of the energy carried away by photons whose wavelengths fall in some given range.

The result of all this is that we shall characterize the spectrum as the *radiated energy flux between two wavelengths*.

Let us consider, then, two wavelengths λ and $\lambda + \Delta\lambda$, where $\Delta\lambda$ is meant to be very small. If $\Delta\lambda$ is small enough, the flux that comes out in this range will simply be proportional to $\Delta\lambda$: if we take half the wavelength range, the energy coming out will be half. (This only works if the spectrum is essentially constant within the range, which will always be true for a small enough range.) We shall call this flux per unit wavelength F_λ :

$$\text{flux between } \lambda \text{ and } \lambda + \Delta\lambda = F_\lambda \Delta\lambda.$$

So the reason we think the two ideas are contradictory is that most everyday objects are simply not hot enough to be bright and black at the same time. But stars are. See Investigation 10.1 for more details.

The color of a black body

Now, the color of the glow of a hot body depends on its temperature. Bodies hotter than the electric burner may glow white hot. Cool bodies, such as the stove’s resting burner, emit radiation too, but we don’t see it because it is in the infrared region of the spectrum, to which our eyes are not sensitive.

Nineteenth century physicists found experimentally that the color or spectrum of the glow emitted by a “perfect” black body depends *only* on how hot it is, independently of what the body is made of or what shape it has. (An example of this spectrum is given in Figure 10.1 on the next page.) Physicists of the time were able to explain satisfactorily why the spectrum depended only on the temperature, essentially by showing that if two black bodies of the same temperature emitted different kinds of radiation and yet absorbed everything (because they were black), then one could use them to construct a **perpetual motion** machine. This fascinating style of argument is common in the branch of physics called **thermodynamics**, but we would be going too far from the theme of this book if we tried to give its details here.

The thermodynamic arguments also incidentally proved that a black body is *more efficient* at giving off light than a body of any other color: this is why, in fact, stove burners are manufactured black. But one thing the nineteenth-century physicists could not explain was the *shape* of the spectrum, or in other words the

This spectrum F_λ for a black body is illustrated in Figure 10.1 on the next page for a few temperatures. Its shape was known from experiment, but nineteenth century physicists could explain only the falling part of the spectrum at long wavelengths. The fact that the curve reached a maximum and turned over was first explained by Planck, who also derived from his arguments on quantized energy levels (see the main text) the famous formula that describes the curve:

$$F_\lambda(T) = \frac{2\pi hc^2}{\lambda^5} \frac{1}{(e^{hc/\lambda kT} - 1)}, \quad (10.1)$$

where T is the temperature, c is the speed of light, k is Boltzmann’s constant (see Chapter 7), and $h = 6.626 \times 10^{-34}$ J s is *Planck’s constant*, which we met in Chapter 7. For readers who have never encountered it before, the symbol e represents a famous and important number in mathematics. It arises in many problems of calculus, as often as π arises in geometry. Its value is approximately $e = 2.71828$. It is a pure number, just like π , and so it carries no units. Raising e to a power, say x , is such a common operation in some parts of mathematics that it is given a special name: $e^x = \exp(x)$. This is called the **exponential function**, and it can be found on scientific calculators (where it is usually called e^x) and as a built-in function in computer languages like Java (where it is called `exp`). It may be used mathematically just like the functions `sin` or `cos`.

It is clear from the curves in Figure 10.1 that the brightness of a body depends on the wavelength and temperature. The body at 5000 K is brightest in the optical part of the spectrum. The body at 100 K is brightest in the infrared. And the body at 10⁶ K is brightest in the X-ray region. The color of the body will clearly depend on its temperature: the cooler the body, the longer the wavelength of most of the light it emits.

Since almost all of our understanding of the Universe can be traced back to the Planck function, we should take some time to understand its properties. This is the subject of Investigation 10.2 on page 117.

In this section: Max Planck founded quantum physics by explaining the spectrum of light emitted by a black body. He postulated that light was always emitted with an energy proportional to its frequency. The proportionality constant, called Planck’s constant h , is one of the most fundamental numbers of physics.

►A perpetual motion machine is a machine that will run by itself forever, without requiring any supply of energy. Since real machines always have a little friction or other losses, perpetual motion requires the creation of energy from nothing. Since this is not possible, a circumstance that would lead to perpetual motion is also not possible.

Figure 10.1. The black-body spectrum for the three temperatures 100 K, 5000 K, and 10⁶ K. Notice how the wavelength at which the maximum occurs decreases with increasing temperature. The function plotted is the energy emitted by the body per unit surface area, per unit time, and per unit wavelength. This is defined in Investigation 10.1 on the previous page.

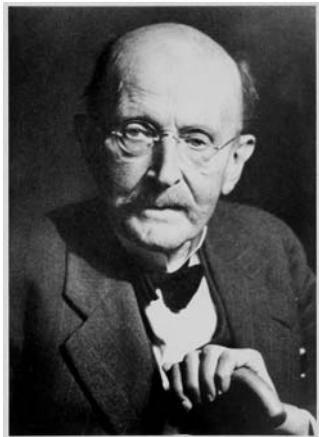
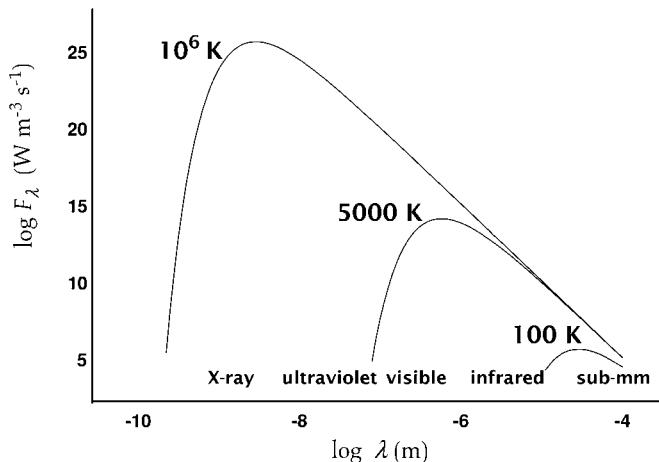


Figure 10.2. Max Planck was one of the leading physicists of the early 20th century. A pioneer of quantum theory, he used his considerable influence in German science to nurture the development of both quantum mechanics and relativity. Near the end of his life he tried in vain to moderate Hitler's attacks on Jewish scientists, including Einstein. He lived to see one of his sons executed by Hitler for treason. Germany's network of pure-science research institutes is now named after this complex and profoundly influential scientist.

Photo courtesy Mary Evans/Weimar Archive.

actual color associated with any given temperature. What is so special about the wavelengths where the curves in Figure 10.1 turn over? Why should a given temperature determine a certain wavelength of light?

The explanation finally came, just at the turn of the century, from Max Planck. With his explanation he made the first tentative step toward quantum theory. Since all previous attempts to explain the spectrum had failed, it was almost inevitable that any explanation would involve a new and strange hypothesis about matter. Planck's new hypothesis was strange indeed. Many physicists believed that the atoms of a black body (or of any other body) emitted light by vibrating. A vibration with a frequency f would emit light at frequency f . When an atom absorbed this light, it would be set into vibration with frequency f . So equilibrium between the radiation and the walls of the black body involved countless events in which light energy was interchanged with vibration energy.

Planck postulated that all the atoms vibrating with a given frequency f could exchange energy only in discrete amounts, only in energy "parcels" of size then its energy of vibration could have only certain values, namely:

$$E = hf, \quad (10.2)$$

where $h = 6.626 \times 10^{-34}$ J s is *Planck's constant*, which we met in Chapter 7.

This is the same formula that Einstein used to explain the photoelectric effect, as we saw in Chapter 8. Einstein worked after Planck, and he gave this formula a more radical interpretation: he assumed that the reason that energy exchanges had to involve only quanta of energy of this size is that light itself can carry only such quanta of energy. For Planck, light was still a continuous electromagnetic wave, and the quantization was something to do with the way atoms behaved. Einstein put the focus onto light itself.

Planck had no theory from which he could predict the value of h . However, when combined with Boltzmann's statistical mechanics of the atoms, Planck's new hypothesis led exactly to the prediction of the shape of the curves in Figure 10.1. Planck could measure the value of h by finding the value that best made his theoretical curve fit experimental measurements, such as those in Figure 10.1. Then, once he had obtained the value of h from the curve for one temperature, he found that he could exactly predict the measurements for all other temperatures. This was the triumph of his theory, and the sign that the constant h was a new and fundamental physical quantity.

Here is why Planck's hypothesis determines a relation between temperature and the wavelength of light: by introducing a new constant h with dimensions of energy times time, it is possible to start with a temperature T , find from it a "typical" energy kT , and from it deduce a number with the dimensions of time, h/kT . This number can be turned into a wavelength λ by multiplying by the speed of light: $\lambda = hc/kT$. Although it may seem that we have just played a mathematical game devoid of any physical reasoning, we have in fact learned one important thing: if the theory does associate a special wavelength with a temperature T , it will probably be roughly the same size as the number hc/kT , since that is the only number with the dimensions of length that we can find in the theory. It was the absence of a special constant like h in the theories of physics before Planck that prevented physicists from finding any special wavelengths associated with light.

Why does quantization of light energy lead to the black-body formula? Although the details are well beyond our scope here, the outline is not hard to grasp. Suppose we are given a hollow box whose rough interior walls are at a temperature T . A tiny hole in the wall of the box will be a black body: shine any light onto the hole and it will go in, with almost no probability of its re-emerging from the hole directly. But the radiation that comes out of the hole will be the same as that inside the cavity, so the cavity contains black-body radiation.

Then the typical vibration energy of the atoms in the walls will be about kT . Changes in these energies can occur only in multiples of hf , because (following Einstein) the properties of light force this. Now, since there is only a finite amount of energy in the walls, and this will be shared among all the atoms, there will be very few with very high frequencies of vibration, because this would involve very large energy exchanges, so there will likewise be little light at very short wavelengths. There will also be few atoms vibrating with nearly zero energy, so there will be little light at long wavelengths. The light distribution must therefore peak at some intermediate wavelength, and that will be proportional to hc/kT . Although Planck's argument went somewhat differently, since he did not then have the benefit of the insight of Einstein, it nevertheless led to the same result: a curve that fit the experimental observations perfectly.

Planck was the first to suggest that energies might in some way be quantized. Planck's hypothesis, bold as it was, was carefully restricted only to the energies of the atoms. We now know that all measurable quantities are quantized: the angular momentum of a spinning particle, the linear momentum of a particle moving around in a box, and so on. This is the province of the quantum theory.

Relation between color and temperature: greenhouses again

How does this relate to stars? A star is black because it absorbs essentially all the light that strikes it. So the spectrum of the light it emits will depend only on its temperature, and this will be the same spectrum as is emitted by hot bodies in the laboratory. In turn, the spectrum uniquely determines the color of the black body, so that hotter bodies emit more energy in the blue, while cooler ones emit more in the yellow. Measuring the color allows one to determine the temperature. In this way we know that stars have surface temperatures ranging from about 2000 K to about 30 000 K.

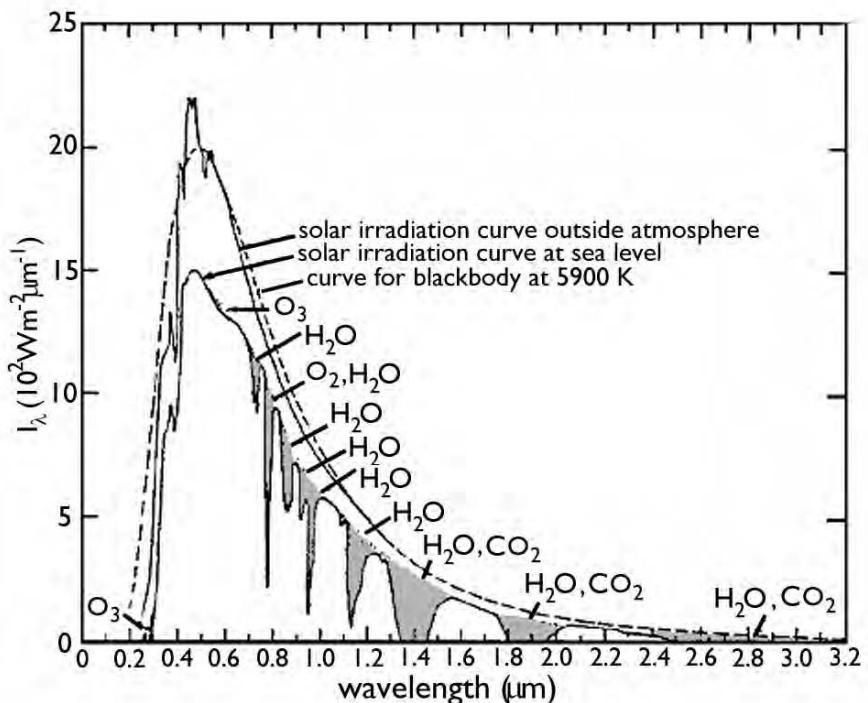
Not only stars, of course, but any body emits radiation with a color or characteristic wavelength that depends on its temperature. This explains the greenhouse effect that we met in Chapter 7. The energy arriving on the Earth from the Sun comes with visible wavelengths, since the temperature of the Sun is about 5000 K. The Earth, with a temperature of only 300 K, re-radiates this energy at much longer wavelengths, in the infrared. This means that it is possible for a gas to allow solar

►This reasoning is another example of dimensional analysis, which we first used in Investigation 1.3 on page 5. Used properly, it can be a powerful first step towards understanding difficult problems.

In this section: greenhouse gases trap heat on the Earth because the Earth is colder than the Sun, so the spectrum of light it emits is different from the one it absorbs.

Figure 10.3. The solar spectrum at the top of the atmosphere is similar to the black-body curve for 5900 K.

The solar spectrum reaching the ground is also illustrated, to show how much of it is absorbed by the atmosphere. The molecules responsible for the absorption are indicated. Figure based on illustration in the CEOS CD-ROM, (<http://ceos.cnes.fr:8100/-cdrom-98/astart.htm>).



radiation to hit the Earth (i.e. to be transparent at visible wavelengths) and still to block radiation leaving the Earth (to be opaque at infrared wavelengths). Greenhouse gases do just this.

Spectral lines: the fingerprint of a star

In this section: the spectrum of light from a star contains features, called lines, that arise in the photosphere, the layer from which photons leave a star. The spectrum contains detailed information about the star itself.

But does this not contradict another thing that we mentioned earlier, namely that the spectrum of a star is closely related to its other properties, such as its mass and size? Yes, in fact, there is a contradiction if we take the black body model too absolutely. In fact, every star is almost, but not quite, black. One reason is that, although stars are mainly made of a hot gas of individual protons and electrons, they contain other elements. If an incoming photon with just the right energy (frequency) strikes an atom of one of these elements in the outermost layers of a star, it can be reflected out of the star. So elements prevent stars from being perfect black bodies. Instead, one finds that the spectrum of a star has an overall shape similar to the black-body curve, but superimposed on this shape are narrow features, called **spectral lines**, that are caused by the absorption or emission of light of particular wavelengths by the atoms in the outer parts of the star. It is these features that are unique to each type of star, and indeed, if we go to enough detail, unique to each individual star.

In Figure 10.3 we can see what the spectrum of the Sun looks like, compared to the spectrum of a black body of 5900 K, which seems to be the temperature that gives the best approximation to the spectrum. The general shape follows the spectrum fairly well, but the spectral lines make noticeable diversions. If we see a star somewhere else whose detailed spectrum matches this, then we may be quite sure that it is similar in size, mass, and composition to our Sun.

Readers who remember how we modeled the Sun in Chapter 8 may wonder

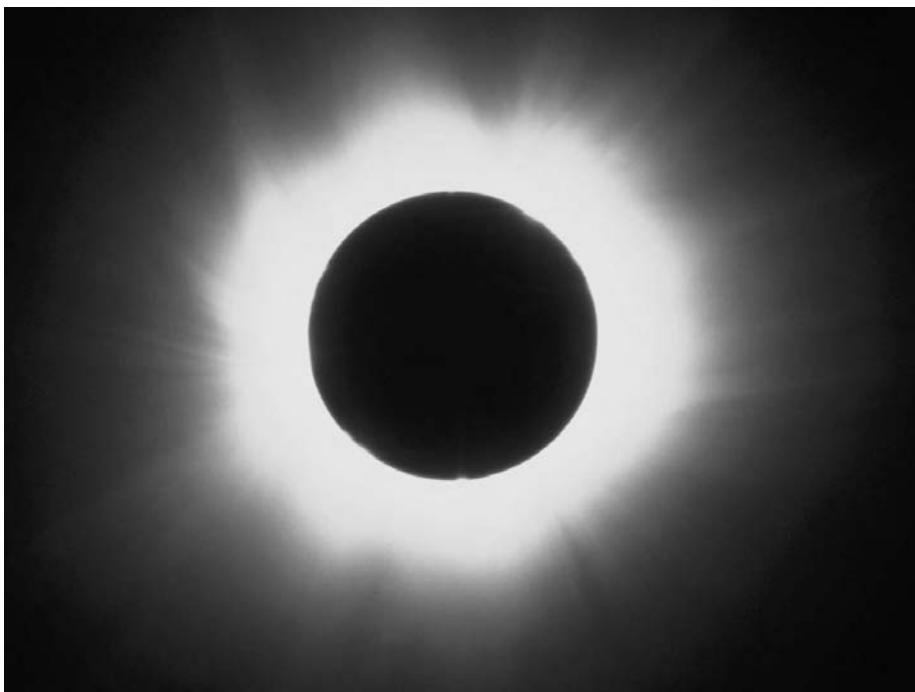


Figure 10.4. The high-temperature corona of the Sun can be seen only when the dominant light from the photosphere is blocked out, such as in this photograph of the 1999 eclipse, taken in Hungary.
(Copyright Pavel Cagas, Zlin Astronomical Society.)

what we mean here by the surface temperature of the Sun. After all, the Sun's temperature drops rapidly near the surface from very large values inside to almost zero, and then rises again to millions of degrees in the very outer regions. Where does the number 5900 K fit into this? The key to understanding this is to recall that photons scatter an enormous number of times as they make their way out from the central regions to the surface. At some point, they become free: the probability of a further scattering before they leave the Sun becomes very small.

These photons are the ones we see with telescopes. Because of the huge number of scatterings that a photon experiences on its way out from the center of the Sun, the region in which this last scattering takes place is localized to a very thin shell, and we call this shell the photosphere. The temperature of the gas at the photosphere is the surface temperature of the Sun. The higher temperatures outside this do not affect most of the light leaving the photosphere, because there is so little gas outside this point that few photons ever get scattered by this gas. It is only visible during an eclipse, when the photons of the photosphere are blocked out (see Figure 10.4).

Notice that all the spectral lines of the Sun in Figure 10.3 are dips in the spectrum rather than rises. This means that light from inside the Sun is being absorbed by the elements responsible for the lines. We call this an **absorption spectrum**. There are occasions, especially in the more exotic objects like quasars (Chapter 14), where spectral lines are seen in emission: there is more light coming to us from the lines than from the black body background.

Why are there lines in the first place? The reason is the modern version of Planck's great insight: atoms can exist in only certain energy states. When they emit or absorb a photon they can make a transition only to another of the allowed energy states, so the photon can have an energy that must similarly follow Equation 10.2 on page 112. In the gas that forms the black body, all vibration frequencies are represented, so the photons come with all wavelengths. But if there are traces of specific elements, which have vibration frequencies characteristic of that element,

Table 10.2. Radii of the five brightest stars, inferred from measured luminosities and temperatures.

Rank	Name	Luminosity (W)	Temperature (K)	Radius (m)	R/R_\odot
1	Sirius	8.0×10^{27}	8000	1.6×10^9	2.3
2	Canopus	7.3×10^{31}	15 000	4.4×10^{10}	63
3	Rigel Kentaurus	5.0×10^{26}	6000	7.1×10^8	1.02
4	Arcturus	3.5×10^{28}	4470	1.1×10^{10}	16
5	Vega	1.8×10^{28}	9500	1.7×10^9	2.5

then they can absorb preferentially at certain wavelengths and produce distinct features. This is what happens to give stars their unique fingerprints.

Now, Planck postulated that the allowed energy states were evenly spaced, the difference in energy being the same from each one to the next. But this does not fit the spectra physicists observe. The next great step toward quantum theory after the work of Planck and Einstein was taken by the Danish physicist Niels Bohr (1885–1962). He devised a more complicated rule in which the spacing in energy decreased as the energy went up, and this rule agreed with the simplest spectra, such as that of hydrogen. Further refinements to the rule allowed it to match more complicated spectra. The different spacing of the energy levels did not undermine Planck’s derivation of the black-body spectrum, because Einstein had already shown that the black-body spectrum only needed the allowed energies of *photons* to be evenly spaced, and not those of the atoms.

How big stars are: color and distance tell us the size

In this section: the Stefan–Boltzmann law says that the luminosity of a star is proportional to the fourth power of its temperature and to its surface area.

From this we find that stars range in size from hundreds of times the size of the Sun down to smaller than the Earth itself.

We come now to one of the most interesting consequences of being able to measure the distance to stars, which is being able to say how big they are. How is it that we can say with confidence that stars range in size from many times the size of the Earth’s orbit down to sizes much smaller than the Earth itself?

The key lies in the black-body spectrum again. The spectrum (color) of the light emitted by a black body depends only on its temperature. But the *amount* of light emitted depends also on the size of the black body: the total light emitted by the black body is just the sum of the light emitted by each patch, and it must therefore be proportional to the *surface area* of the black body.

This is a remarkably simple conclusion: if we measure in the laboratory the total emission of a black body at some temperature T , and then if we find another black body of the same temperature that emits twice as much light, it must have twice the area. For stars, which are basically spherical in shape, if we know the area then we know the radius. We described earlier how knowing the distance to a star tells us the total energy emitted by it (its absolute magnitude). We also saw that measuring the color of the star tells us its temperature. These two together then tell us how big the star is.

In Investigation 10.2 we shall see that the law relating luminosity, area, and temperature is remarkably simple. The luminosity is proportional to the area and to the fourth power of the temperature:

$$L = \sigma AT^4, \quad (10.3)$$

where σ is the constant of proportionality called the *Stefan–Boltzmann constant*. (Yes, Boltzmann turns up here too!) Its value is

$$\sigma = \frac{2\pi^5 k^4}{15c^2 h^3} = 5.67 \times 10^{-8} \text{ W m}^{-2} \text{ K}^{-4}, \quad (10.4)$$

where the units are watts per square meter of surface area per degree kelvin to the fourth power. This is called the Stefan–Boltzmann law.

Investigation 10.2. Exploring the Planck Function

There are two very important properties of the Planck spectrum that are not hard to understand, but which are central to the way astronomers use the function to learn more about stars.

- The first property is that the wavelength where the peak of the spectrum occurs is inversely proportional to temperature. (This is called Wien's law.)
- The second is that the total luminosity of a black body is proportional to the fourth power of its temperature. (This is called the Stefan-Boltzmann law.)

In fact, both laws were known to physicists before Planck deduced the full theory of thermal radiation, but it is easier for us to understand them now as a consequence of his theory.

Although the expression for F_λ may seem so complicated that these results might be difficult to prove, the situation is actually rather simpler than that. The key to simplifying F_λ is to give a simple name to the exponent of e in Equation 10.1 on page 111. We define a new variable x to be:

$$x = hc/\lambda kT. \quad (10.5)$$

The various quantities in this expression all have complicated units of measurement (especially h), but we now show that the units must all cancel out to give a number with *no* units at all, a **dimensionless number**. This is because x enters F_λ as an exponent, a power: Equation 10.1 on page 111 contains

$$e^{hc/\lambda kT} = e^x.$$

Now, one can raise a number, say 3, to a power that is a pure number, like 5, to get $3^5 = 243$; this means that we multiply together five factors of 3. But if the power has dimensions, the expression has no meaning: what are 5 km factors of 3? Interested readers should check that x is indeed dimensionless, but it must work out that way for F_λ to make any sense at all. (Readers may also like to verify that x is just proportional to the *frequency* of the light, in fact that x is just the ratio of the frequency f to what one might call the "thermal frequency" kT/h , which is the frequency a photon would have if its energy were equal to the typical thermal energy kT .)

Let us now take x to be a variable in F_λ instead of λ itself. If we solve Equation 10.5 for λ we get

$$\lambda = hc/xkT. \quad (10.6)$$

We now substitute this into Equation 10.1 on page 111 to get

$$F_\lambda = \frac{2\pi k^5 T^5}{h^4 c^3} \frac{x^5}{e^x - 1}. \quad (10.7)$$

Apart from a coefficient out front that is constant for a given temperature, the function at the heart of this is the dimensionless function

$$f(x) = x^5/(e^x - 1), \quad (10.8)$$

whose properties depend only on x .

Now we can look at the laws we wish to establish, the Wien and Stefan-Boltzmann laws. The peak of the spectrum for a given temperature occurs where $f(x)$ reaches a maximum. Let us call this value x_{\max} . We don't need to know its value. All we need is the relation above for λ as a function of x : the maximum will occur at the value of λ given by

$$\lambda_{\max} = hc/x_{\max} kT.$$

This proves that the peak wavelength is inversely proportional to T . In Investigation 10.3 on page 119 we will see how to calculate the value of x_{\max} . The result gives the Wien law

$$\lambda_{\max} = 0.29/T \text{ cm}, \quad (10.9)$$

where T is given in degrees kelvin. For example, if a spectrum peaks in the visible region, say at 0.5 microns (5×10^{-7} m), then its black-body temperature is 5900 K, just like the Sun.

The Stefan-Boltzmann law has a similar foundation. The energy radiated between λ and $\lambda + \Delta\lambda$ is, for sufficiently small $\Delta\lambda$, just equal to

$$\text{flux} = F_\lambda \Delta\lambda.$$

Now we need to convert $\Delta\lambda$ to an equivalent range of the variable x . The wavelengths λ and $\lambda + \Delta\lambda$ correspond to two values of x . Since x decreases as λ increases, we call these values x and $x - \Delta x$, respectively:

$$x = \frac{hc}{kT} \frac{1}{\lambda}, \quad \text{and}$$

$$x - \Delta x = \frac{hc}{kT} \frac{1}{\lambda + \Delta\lambda}.$$

Their difference gives Δx :

$$\begin{aligned} x - (x - \Delta x) &= \Delta x \\ &= \frac{hc}{kT} \left(\frac{1}{\lambda} - \frac{1}{\lambda + \Delta\lambda} \right), \\ &= \frac{hc}{kT} \left(\frac{\Delta\lambda}{\lambda(\lambda + \Delta\lambda)} \right) \\ &\approx \frac{hc}{kT} \left(\frac{\Delta\lambda}{\lambda^2} \right), \end{aligned}$$

where the last step is an approximation that gets better and better as we make $\Delta\lambda$ smaller and smaller. Next we replace the factor λ^2 with its equivalent in terms of x to get

$$\Delta x = \frac{kT}{hc} x^2 \Delta\lambda.$$

We can now solve this for $\Delta\lambda$ to get

$$\Delta\lambda = \frac{hc}{kTx^2} \Delta x.$$

If we put this into the flux expression and use Equation 10.7 for F_λ , we find

$$\text{flux} = \frac{2\pi k^4 T^4}{h^3 c^2} \frac{x^3}{e^x - 1} \Delta x. \quad (10.10)$$

This is proportional to a pure number depending on x and to T^4 . If we ask for the total flux of light from the body, we have to add up contributions like this from all ranges of wavelengths, from $x = 0$ (the longest wavelengths) to $x = \infty$. Then even the dependence on x goes away, and the result is that the total flux of energy radiated from the surface of a black-body is proportional to T^4 . The full equation for this is justified in Investigation 10.3 on page 119:

$$F = \frac{2\pi^5 k^4}{15c^2 h^3} T^4. \quad (10.11)$$

The dependence on k , T , h , and c is as in Equation 10.10 above. The pure numbers (such as π^5) come from finding the area under the curve $x^3/(e^x - 1)$, which we do in Investigation 10.3 on page 119.

Now we can at last test whether our ideas about black bodies have any relation to the real stars, in particular to the Sun. Each square meter of the surface of a black body at a temperature of 5900 K, like the Sun, shines with a power of 6.42×10^7 W. Since from Equation 9.1 on page 106 the Sun's luminosity is 3.83×10^{26} W, it must have a surface area of $3.83 \times 10^{26}/6.42 \times 10^7 = 5.97 \times 10^{18}$ m². From the formula for the area of a sphere, $A = 4\pi r^2$, it follows that the solar radius is 6.89×10^8 m.

This is very close to the accepted value of 6.96×10^8 m, which is obtained by direct measurements by spacecraft. The closeness of these two numbers is a triumph for the black-body model of the Sun! We can expect to use it with confidence on other stars.

In Table 10.2 on page 116 we look at the same stars as in Table 10.1 on page 110, only this time we list their luminosities and temperatures and the radii we infer by the method we have just applied to the Sun. There is a huge range of sizes, from about the size of the Sun to more than 60 times its size. This small selection of stars illustrates an important point: most stars are either about the same size as the Sun or they are big. Astronomers call the normal stars **main sequence stars** for reasons that will become clear in the next chapter. The big stars are called **giants**.

In fact, the range of size is even greater than we have shown with this selection of stars. For example, the star Sirius is actually two stars, one of which is very dim. Called Sirius B, its luminosity is 9.6×10^{23} W, but its color is very blue and its temperature is a very high 14 500 K. The only way it can have such a high temperature and yet be so dim is for it to be small. The radius we infer is only 5.5×10^6 m, or 0.86 times the radius of the Earth! This star is truly remarkable, because we can see the gravitational effect it has on its much brighter companion, Sirius A, and infer from this that its mass is actually 1.05 times the mass of the Sun. (We will see in Chapter 13 how to do this.) It has more than the mass of the Sun squeezed into a volume smaller than the Earth! Such a star is called a **white dwarf**, and we will find out in Chapter 12 how such extraordinary stars can exist.

Even this is not the end of the scale of sizes. When we come to study pulsars in Chapter 20, we will see that they are neutron stars. They have masses greater than the mass of the Sun, yet their typical radii are only 10 km, smaller than a good-sized city! Because their sizes are inferred by means other than those we have employed here, we shall reserve a full discussion of these incredible objects to the later chapter.

But why are stars as hot as they are, and no hotter?

In this section: the surface temperatures of most stars are not very different, and this comes from the way photons free themselves from the star.

We have come a long way in our understanding of stars just by learning that they are black bodies. We have used that knowledge to measure their sizes. But we can do even better: with a little thought, we can actually predict the temperatures of the stars. Notice a rather remarkable feature of Table 10.2 on page 116. The range of temperatures of the stars in the table is not large. While their radii range over a factor of about 60, and their luminosities over more than 10^5 , their temperatures differ by less than a factor of four. Is there something, then, that fixes the temperature of the star?

To answer this we must remind ourselves that we are looking at the surfaces of the stars, not their interiors. And then we meet a puzzle: when we solved for the structure of the Sun in Chapter 8, we found that we predicted that the temperature of the Sun should fall smoothly to zero at its surface! (In fact, as we noted in Chapter 8, complex dynamical processes heat the outer corona of the Sun to a very high temperature, but so little mass is out there that it has no influence on the normal visible properties of the Sun.) What, then, do we mean by the temperature of a star: is not its surface temperature zero?

We can solve this puzzle by thinking about what happens to photons trying to get out of the star into space, eventually to hit our eye. Photons have a special affinity for charged particles, like electrons and protons. Radiation is given off by any moving, accelerating charge: the radio waves coming from the radio transmitting tower near your home come from electrons that race up and down the tower at the right frequency. Neutral atoms accelerating give off no radiation. The time-reverse of emitting radiation is absorbing it, and the same considerations apply: a light beam

Investigation 10.3. Computing the Planck function

Readers who have access to a computer and who already understand the exponential function e^x and its inverse, the natural logarithm $\ln(x)$, may wish to verify the graphs in Figure 10.1 on page 112 and the results of Investigation 10.2 on page 117 by using the computer program Planck on the website. In order to understand how the program is constructed, we have to look at a difficulty that arises in calculating functions like $f(x)$ that depend on the exponential function e^x .

If one simply programs the expression for $f(x)$ directly, one soon finds a difficulty: the exponent on e can get very large and overflow the limits that ordinary computers set on such things ($|x| < 200$ or so). The way to get around this problem is to calculate the natural logarithm of $f(x)$,

$$\ln[f(x)] = 5 \ln x - \ln(e^x - 1), \quad (10.12)$$

and then to make separate approximations to the exponential function for three different ranges of the variable x . These are as follows.

- If we get into trouble because x is too large an exponent for the computer to be happy with, then of course we will also have $e^x \gg 1$ so we may neglect the 1 subtracted from e^x in the second term of Equation 10.12. This means we have

$$\ln(e^x - 1) \approx \ln(e^x) = x,$$

the last equality following from the inverse property of the exponential and the natural logarithm.

- If x is neither large nor small, say $0.01 < x < 100$, there is no problem evaluating Equation 10.12 directly in the computer, so no approximation is needed.
- If x is smaller than, say, 0.01, we can make use of a remarkable (and very profound) property of the exponential function, namely that for small x we can approximate

$$e^x \approx 1 + x, \quad |x| \ll 1. \quad (10.13)$$

If this seems surprising, remember that any number raised to the power of zero is equal to 1. So it is not surprising

that when x is nearly zero then e^x is nearly 1. What is remarkable is that e^x differs from 1 just by x itself! Experiment with this on your pocket calculator. For example, if $x = 0.1$ I find that $e^x = 1.10517$. This is pretty good: the error of the approximation is 0.005, or one-half of one percent of the answer. The approximation works if x is negative as well, and it gets better as $|x|$ gets smaller. This allows us to make the following approximation in Equation 10.12:

$$\ln(e^x - 1) \approx \ln(1 + x - 1) = \ln x.$$

These three cases are the basis of the program Planck given on the website, which calculates $f(x)$.

The program finds the maximum of $f(x)$ in a simple way: as it steps through its range of values of x , it tests each computed value of $f(x)$ to see if it is larger than the largest previous value. If it is, then this value becomes the new maximum and its value of x the new x_{\max} . After all values of x have been tested, we have found the "global" x_{\max} . The program finds the maximum to be $f_{\max} = 21.20$ at $x_{\max} = 4.95$. This value of the maximum leads to Wien's law, Equation 10.9 on page 117, when we note that $x = hc/\lambda kT$. Solving for λ_{\max} gives

$$\lambda_{\max} = \frac{hc}{x_{\max} kT} = \frac{0.29 \text{ cm}}{T},$$

when T is measured in degrees kelvin. This is just what we quoted in Equation 10.9 on page 117.

The program also finds the coefficient in the Stefan-Boltzmann law by simultaneously computing the area under the curve $x^3/(e^x - 1) = f(x)/x^2$. This is equivalent to summing up the fluxes from all the little wavelength ranges, each given by Equation 10.10 on page 117. The method for estimating the area under a curve using a computer is described in Figure 10.5. The program Planck gives 6.494. More sophisticated calculations using the tools of the calculus give an exact value of $\pi^4/15$. This evaluates to 6.494, keeping the same number of places. Our simple computer program has given us the right answer for Wien's law and the Stefan-Boltzmann law to several places accuracy, and we did not have to use any advanced mathematical techniques.

has an easy time making charged particles move, and thereby being partly absorbed and partly scattered as a result. But light passing through a medium with no free electrons is little affected by it. Indeed, metals are good at reflecting light because the electrons in them are free to move around, whereas in transparent materials like glass the electrons are firmly held to their atoms.

Because the Sun is a highly ionized gas, composed of bare protons and electrons, light has a hard time passing through it. Photons in the hot inner regions of the star do not reach us directly: they scatter off the electrons and protons too readily. However, as one goes outwards from the center of the star to its surface, the temperature of the gas decreases, until eventually it falls low enough to permit hydrogen (its main constituent) to remain in its neutral atomic state. At this point, photons are free to escape and reach us. So on this argument, we would expect the surface temperature of the star to be about the temperature required to ionize hydrogen.

This temperature is not hard to calculate. The energy required to pull an electron

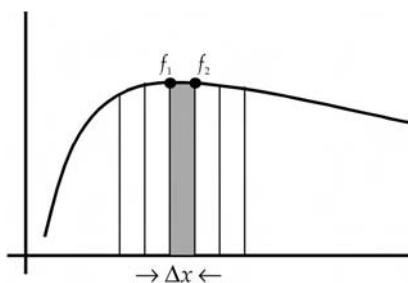


Figure 10.5. For Investigation 10.3, we approximate the area under the curve $f(x)$ by dividing it into many small sections of width Δx and replacing the curve by a straight line across the top of each section. The approximation to the area of the shaded trapezoidal section shown here is then $(f_1 + f_2) \times \Delta x/2$. The approximation is good if Δx is small enough that the curve is practically straight across the section.

>Recall the definition of the electron volt (eV) given in Chapter 8: $1 \text{ eV} = 1.602 \times 10^{-19} \text{ J}$.

away from the proton in a hydrogen atom is about 13 eV. In a hot gas, this is provided by collisions: if the energy transferred by one atom to another in a collision exceeds this amount, the result is likely to be that an electron gets knocked off an atom. The energy transferred in a collision should be comparable to the energy of the moving gas atoms, which is about $\frac{3}{2}kT$. Setting this equal to 13 eV gives a temperature of $T = 10^5$ K. This is about a factor of 2 larger than the Sun's surface temperature, which means that our reasoning is pretty good but not perfect. The main error is simply that we don't need *all* the photons to be energetic enough to ionize the hydrogen atoms. If the temperature were lower, there would still be photons with energies above 13 eV, but there would just be fewer. That would be acceptable if there are still enough of them to ionize any hydrogen atoms that happen to form by the recombination of a proton and electron. Exactly what this temperature should be depends on a balance that involves the density of the gas (which affects how often hydrogen atoms form by recombination) and the details of the distribution of energies of different photons at any given temperature. We need not worry about these details here.

Looking ahead

In this section: we have learned how to determine the overall properties of a star. This forms our platform for investigating what goes on inside in the next chapters.

What we have learned here and in the previous chapter sets the stage for the next four chapters. Now that we have seen how astronomers have learned how far away, how bright, how hot, and how big stars are, it is not difficult to guess that we will be able to discover much more about what goes on inside them. In the next chapter we see how the balance between the inward pull of gravity and the outward pressure of hot gas is maintained by the steady nuclear reactions that make stars shine, and from which life itself ultimately derives. Then we will look at what happens when this balance fails, and stars "die", turning themselves into white dwarfs, neutron stars, or even black holes. After that we will look at stars in pairs, orbiting one another, sometimes so closely that they exchange gas and even feed black holes. Then in Chapter 14 we will look at the ways that stars form larger groups held together again by gravity: star clusters, galaxies, clusters of galaxies. We shall again have to calibrate our rulers and change our perspective on size: the distances between galaxies will make the distances between stars seem minuscule.

Our study of galaxies will conclude the first half of the book. After that, we have to widen our ideas about fundamental physics to include relativity. Once we do that, the second half of the book opens up to us: black holes, gravitational waves, cosmology, quantum gravity, and the beginning of time itself.

Stars at work: factories for the Universe

In this chapter we open the door to our own history. Surely one of the most satisfying discoveries of modern astronomy is how the natural processes of the Universe led to the conditions in which a small planet could condense around an obscure star in an ordinary looking galaxy, and life could evolve on that planet.

The evolution of life seems to have required many keys, but one of them is that the basic building blocks had to be there: carbon, oxygen, calcium, nitrogen, and all the other elements of living matter. The Universe did not start out with these elements. The Big Bang, which we shall learn more about in Chapter 24 to Chapter 27, gave us only hydrogen and helium, the two lightest elements. All the rest were made by the stars. Every atom of oxygen in our bodies was made in a star. It was then expelled from that star and eventually found its way into the hydrogen cloud that condensed into the Sun and the Solar System.

Indeed, the atoms in our bodies have a durability that makes the time they spend being part of us seem minute. All the elements participate in vast recycling schemes on Earth: the carbon dioxide cycle (which we mentioned in Chapter 7), the calcium cycle, and so on, in which they move in and out of living organisms. But these cycles are mere epicycles on a grander cycle, in which the atoms are made in stars, pushed out into giant interstellar clouds of gas, and incorporated into new stars. In their new stars they might actually be torn apart and the pieces – neutrons and protons – re-assembled into new elements. They might then be expelled to go around the grand cycle once more, or they might become trapped in a dying star and be lost forever.

This grand cycle of stellar birth, death, and re-birth is the main business of the Universe. The balance between gravity and nuclear reactions drives this cycle. We shall look more closely at the processes of birth and death in the next chapter. Here we look at what keeps the star going during its normal lifetime.

Star light, star bright ...

If we want to understand what the stars are doing, let us go straight to their centers and ask: where does the energy come from that makes the stars shine? The short answer is: nuclear reactions. The conversion of hydrogen and helium into other elements gives off energy, and that comes out as light from the star's surface. Physicists only learned about nuclear physics in the twentieth century. Nineteenth century astronomers speculated about the subject, but they did not have enough understanding of physics to know what made the stars shine. We shall first see why the mechanisms they knew about could not work, and then see why nuclear reactions do.

If you are an astronomer speculating about why the stars shine, but you don't know about nuclear physics or you want to look for other sources for the energy radiated by the Sun and stars, then two possibilities should come to mind. One is that perhaps there are some chemical reactions going on inside. After all, chemical reactions are the source of most of our heat on Earth: burning wood, gas or coal

In this chapter: we look at the way stars have created the chemical elements out of which the Earth, and our bodies, are formed. The nuclear reactions in generations of stars that burned out before our Sun was formed produced these elements. But the physics is subtle, and nearly does not allow it. We examine this issue, and also show how the study of a by-product of nuclear energy generation in the Sun, neutrinos, has revealed new fundamental physics.

In this section: we ask where the energy of the Sun could be coming from. Chemical reactions or gravitational contraction cannot supply enough energy. Nuclear reactions can, because they convert mass into energy. Moreover, they have a characteristic signature: the Sun should be emitting the elusive particles called neutrinos.

releases energy that heats our houses and, after conversion into electricity, lights our rooms. Chemical reactions differ from nuclear ones in that they do not convert one kind of atom into another; they just rearrange the ways that atoms combine to form molecules. Maybe stars are forming certain kinds of molecules and releasing heat that way.

The other possibility is that stars are contracting, getting smaller, falling gradually inwards upon themselves. This would release energy. This energy is the same sort that would be released if a small mass fell to the Earth into, say, a bucket of honey: the energy of its fall is dissipated by friction in the honey, and the result is that the honey is heated slightly. If enough energy were dissipated this way, the honey would be hot enough to shine (if it didn't boil away first!).

The problem with both of these explanations is that neither chemical nor contraction energy can last long enough. Let us take the Sun in particular, since it is the star we know most about. The Earth is known to be about 4.54 billion years old, from studies of the radioactive decay of elements in its rocks, and the geological evidence is that the Sun has had pretty much the same luminosity for all that time. We show in Investigation 11.1 how to calculate how much energy could come from either process.

Our calculation shows that the Sun could not derive its luminosity from gravitational energy for more than 1% of its present age, and that chemical energy is even less effective. Gravitational contraction could not have released enough energy to have kept the Sun shining for more than a few tens of millions of years. Nineteenth century astronomers nevertheless believed that contraction powered the Sun, and this led them into great conflicts with geologists, who knew that the Earth was older, and with Darwinian evolutionists, who needed much longer time-scales for evolution to work.

The puzzle of the energy source for stars began to clarify the moment Einstein discovered his famous equation $E = mc^2$, about which we will have much more to say in Chapter 15. For now we only need to know that Einstein predicted that mass and energy can be converted into each other. We see in Investigation 11.2 on page 124 that, if one could find a process that converts even a small fraction of the mass of a hydrogen atom into its equivalent energy, there would easily be enough energy to power the Sun. However, physicists did not know what the details of such a process might be until the discovery of the neutron in the 1930s.

The key advance was the realization that one could *make* helium out of hydrogen by a nuclear reaction that essentially converts four protons (which are the nuclei of hydrogen) and two electrons into a single nucleus of helium, which consists of two protons and two neutrons. This nucleus is also called an alpha particle. This reaction gives off energy, converting about 0.8% of the mass of each hydrogen atom into energy. As Investigation 11.2 shows, this is more than enough to power the Sun for the required time. The reaction does not happen easily, though. Nuclear reactions only occur if particles get very close, and this does not happen often because the protons have electric charges that repel each other. To get them close enough together, they must have a large enough speed to overcome the electric repulsion. In stars, they get this speed from the random thermal motion of gas particles at the temperature of the core of the star. This means that nuclear reactions only occur if the core is hot and dense enough. They do not occur in the core of the Earth, for example. The details of the reactions involved are explored in Investigation 11.4 on page 127.

Investigation 11.1. What doesn't make the stars shine

Since we know the Sun is at least 4.54 billion years old, we can ask whether any proposed source of energy could last that long. Suppose every hydrogen atom in the Sun did something once during the lifetime of the Sun that released energy: it engaged in a chemical reaction, or it simply fell into the Sun and released its energy of fall. How much energy would it have to release on that occasion to be able to account for its share of the steady luminosity of the Sun?

The Sun has a mass of $M_{\odot} = 2 \times 10^{30}$ kg. Each hydrogen atom has a mass of $m_p = 1.67 \times 10^{-27}$ kg. Therefore there are $N_p = M_{\odot}/m_p = 1.2 \times 10^{57}$ protons (hydrogen atoms) in the Sun. (The electrons are so light that they don't affect this calculation. We are ignoring the other elements, since we will see that errors of 10 or 20% won't be important to our conclusions.) Geologists find evidence that the Earth has had liquid water (oceans) on its surface for its whole history, so we can conclude that the Sun has been shining with roughly its present luminosity of $L_{\odot} = 3.8 \times 10^{26}$ W for at least $t_{\odot} = 4.5 \times 10^9$ y = 1.4×10^{17} s. Therefore it has given off a total energy of $E_{\odot} = L_{\odot} \times t_{\odot} = 5.4 \times 10^{43}$ J so far. This is an energy of $E_{\odot}/N_p = 4.5 \times 10^{-14}$ J per hydrogen atom. Let us call this the "energy duty" on atoms in the Sun: in order to join the Sun, the average hydrogen atom is required to "pay" on average 5×10^{-14} J once during its time in the Sun.

Now what of our two candidate sources of energy – are they capable of paying this duty?

- **Chemical energy.** There are so many possible chemical reactions that one might think it would be impossible to decide whether chemical reactions could do the job. After all, could there not be some mysterious chemical at work whose reactions give off enormous amounts of energy? The answer is no: all chemical reactions involve separating electrons from atoms and placing them around other atoms, and so chemical reactions involving any given atom cannot give off more

energy than it would take to remove an electron completely from that atom. This energy is called the ionization energy of the atom, and for all atoms it is less than or about equal to 10 eV. (Recall the definition of the electron volt (eV) given in Chapter 8: 1 eV = 1.602×10^{-19} J.) Thus, an atom of hydrogen engaged in a chemical reaction could only pay an "energy duty" of about 10^{-18} J, too small by a factor of more than 10^4 . (Multiple chemical reactions are no solution: once the electron has been given up, it can't be given up again, and hydrogen has only one to give.) Chemistry is not the answer: it could make the Sun shine for perhaps a million years, not ten billion.

- **Gravitational energy of infall.** The Sun was formed from a cloud of gas that contracted, heating itself up. Could the Sun still be contracting and generating heat? The energy it would generate per hydrogen atom would be roughly the same as one would get if one allowed an atom of hydrogen to fall onto the present Sun from far away. (Again, errors in our estimates of factors of 2 or even 10 will turn out not to matter to our conclusions.) This is just the reverse of "launching" an atom from the surface of the Sun and letting it escape from the Sun's gravity. The speed the infalling particle would have when it hit the Sun is then exactly the same as the escape velocity of the Sun. In Chapter 4 we saw that the escape velocity from the Sun is $v_{\text{escape}} = (2GM_{\odot}/R_{\odot})^{1/2}$. The kinetic energy of the atom when it reaches the Sun will thus be $\frac{1}{2}m_p v_{\text{escape}}^2 = GM_{\odot} m_p / R_{\odot} = 3 \times 10^{-16}$ J. This is better than for chemical reactions, but still inadequate: the Sun might shine by contraction for 100 million years, but not 5 or 10 billion.

Some other energy source is needed. Only nuclear reactions seem to be able to do the job.

Exercise 11.1.1: Chemical bangs

Can it really be true that chemical reactions all give off the same energy per atom, to within a factor of, say, 10 or 100? Don't the chemical reactions that make a TNT bomb explode give off far more energy than the chemical reactions that heat up a smoldering rubbish dump? Explain why the answer to this question is no.

Exercise 11.1.2: Turning on the lights

Once a cloud of gas begins to contract to form a star, roughly how long does it take before nuclear reactions begin to power the star? Will the star shine before this?

The conversion of hydrogen to helium is not quite as simple as suggested in the last paragraph. It turns out that one of the particles produced as a by-product of the reaction is the neutrino. One of Nature's most elusive particles, the existence of the neutrino was guessed at in the 1930s, by theoretical arguments that we will describe in Chapter 16. But it was not directly detected until experiments during the period 1953–1956, by the American physicists Clyde Cowan (1919–1974) and Frederick Reines (1918–1998).

Despite its elusiveness, the neutrino plays a key role in many problems in astronomy, in particular powering supernova explosions (Chapter 12). Neutrino astronomy is in its infancy, but we will see below that it has already completely revolutionized our understanding of the physics of neutrinos. One reason for the neutrino's importance is that it is emitted in abundance whenever nuclear reactions occur in astronomy. For example, two neutrinos are released every time a helium nucleus is created. Another reason is its very elusiveness: it travels fast (nearly at the speed of light), and it hardly interacts with anything at all. This means it can go through matter, such as the outer layers of a star, with little hindrance. Only gravitational waves, which we will meet in Chapter 22, have greater penetrating power than neutrinos.

>This confirmation of the neutrino won belatedly for Reines a share of the 1995 Nobel Prize for physics, but this honor came too late for Cowan, who had already died.

Investigation 11.2. What does make the stars shine

Mass and energy are related. We will see in Chapter 15 that relativity tells us that if a system gives off an amount of energy E , then its mass must decrease by E/c^2 , where c is the speed of light. For chemical reactions, the release of, say, 1 eV from a reaction involving, say, a pair of oxygen atoms makes a negligible change in the mass of the system. The lost mass is $1 \text{ eV}/c^2 = 2 \times 10^{-36} \text{ kg}$, while the masses of the atoms together amount to about 32 proton masses, or $5 \times 10^{-26} \text{ kg}$. So the system loses less than one part in 10^{10} of its mass in a chemical reaction.

When nuclear reactions take place, on the other hand, the mass changes are significant, so that we can deduce the energy released just by looking at the masses of the particles involved. If four protons combine to form a helium nucleus, the starting mass is $6.69 \times 10^{-27} \text{ kg}$, while the final mass is $6.64 \times 10^{-27} \text{ kg}$, a change of $5 \times 10^{-29} \text{ kg}$, or almost 1%. This is more than 10^7 times as much energy as is released in a chemical reaction. (Other particles, usually

electrons or **positrons**, are also involved in these reactions in order to keep the total electric charge the same before and after, but the electron mass of $9 \times 10^{-31} \text{ kg}$ is negligible here.) In energy units, it amounts to $4.5 \times 10^{-12} \text{ J}$ released for every helium nucleus formed.

By comparison with the energy duty of $5 \times 10^{-14} \text{ J}$ that we saw in Investigation 11.1 was required of each proton in the Sun, this is huge. Nuclear reactions have more than enough energy to power the Sun. In fact, is this excess too much of a good thing? Does it mean that the Sun ought to be either more luminous or more long-lived than it is? The answer is no: the nuclear reactions only take place at the very center of the star, where the temperature and density are high enough. So only about 1% of the mass of the star actually gets to be involved in the nuclear reactions that convert hydrogen to helium. We will see in Chapter 12 what happens when these reactions have exhausted the hydrogen fuel in the central part of the star.

Exercise 11.2.1: Water power

If all the hydrogen in a teaspoonful of water were converted into helium, how long would that water power a 100 W light bulb? Take a teaspoon to contain 5 g of water.

...first star I see tonight

In this section: what does the emission of light by the Sun and other stars tell us about the history of the Universe itself? We argue that the Universe cannot be infinitely old: it must have had a beginning.



Figure 11.1. Fred Hoyle was one of the most influential astrophysicists of the twentieth century. We will discuss his fundamental contributions to the understanding of how elements are made in stars below. He is also credited with having coined the phrase "Big Bang" for the beginning of the Universe. He seemed to relish taking controversial scientific positions, of which the C-field was an example. In his later years he became embroiled in heated debates about panspermia, also mentioned below. (Reproduced courtesy of N C Wickramasinghe.)

Let us step back from the details of the source of the Sun's energy to think for a moment about where our investigation is leading us. The Sun began its life as a star roughly 5 billion years ago. What was the Universe like before that? Were there other stars, which were formed earlier, and may by now have died? If so, then what was the Universe like before them? Another generation of stars, perhaps? But earlier again – could the process of forming stars have been going on forever? Is the Universe infinitely old?

The answer is that it can't be infinitely old. In any region of space, say the space occupied by our Galaxy, there is a finite amount of hydrogen. Every generation of stars converts some amount of it into helium. If this had been going on for too long in the past, we would not have much hydrogen in stars today. Yet all stars we see are primarily composed of hydrogen. Observations leave room for maybe three or four earlier generations of stars, but not for an infinite number.

Could there be a way out of this: maybe the helium is somehow re-converted back into hydrogen by some mysterious process? Any such process would need to replace the energy that the original reaction gave off, and where would the process get such energy? The original energy has simply gone away, carried into empty space by the light given off by the stars. Any energy for making hydrogen again would have to come from other nuclear reactions, and they would in turn lead to the same argument.

Time marches on. The Universe *does* get older. Conversely, there was a time when the Universe was young. There *was* a first star.

It is important to understand that this is not in itself an argument for what we call the "Big Bang", although it is a step toward it. The Big Bang is now the almost universally accepted model for how the Universe began. Our arguments above from nuclear physics do not tell us how the Universe began. All they tell us is that, before a certain time not too long ago (as measured in stellar lifetimes), the Universe was doing essentially nothing. The theory of the Big Bang makes a further step beyond this to say that the Universe expanded from a point a few stellar lifetimes ago, and it probably didn't exist at all before that time. We will look at the reasons we believe this, and what it actually means, in Chapter 24.

The only way to avoid the conclusion that the Universe has a finite age is to postulate that somehow energy is simply created in order to replace that which is lost as stars shine. The British astrophysicist Sir Fred Hoyle (1915–2001) and the Indian astrophysicist Jayant Narlikar (b. 1938) made such a postulate, called the **C-field** (“C” for “creation”). They invented this field in order to support the so-called **Steady-State model of the Universe**.

The C-field makes matter to fill in the empty spaces of the Universe left behind as it expands, so that it can be “steady” while expanding. We will return to this subject in Chapter 24. The weight of observational evidence today is heavily against the Steady-State model and its C-field.

Cooking up the elements

Once we accept that nuclear reactions are changing hydrogen into helium, it is natural to ask what other reactions are happening. In particular, if the amount of helium in the Universe is increasing with time, is that also happening to other elements?

The answer is yes: just as the lightest element, hydrogen, can make helium, so can several helium nuclei react to form heavier elements, such as carbon, oxygen, and silicon.

Much of the story of the formation of the elements depends, of course, on the details of the nuclear physics, but one fact is of overriding importance: the most tightly bound collection of protons and neutrons that it is possible to form is the nucleus of iron.

This means that the sequence of reactions forming elements heavier than helium stops with iron. If iron reacts with anything else to form a heavier nucleus, energy must be added to make the reaction go.

We do of course find other elements on the Earth. One of the most important heavy elements is uranium, more than four times heavier than iron. But there is not much of it around, and that means it was formed in unusual circumstances, when energy from some source other than nuclear reactions was available to drive the reactions past iron. The fact that uranium is formed by the addition of energy to other nuclei partly explains why it is good as a nuclear fuel: by splitting it apart into smaller nuclei, one liberates some of the energy that was put in originally.

We believe that most uranium was formed in the giant stellar explosions that we call *supernovae*. We shall explain in the next chapter what supernova explosions are and where the extra energy comes from that goes into making uranium and other heavy elements. It is interesting to reflect on the fact that the ultimate source of our nuclear energy on Earth is not our Sun.

The Sun is responsible for most of the forms of energy we use, such as chemical (burning coal, oil, gas, or wood) or environmental (wind, hydroelectric, and direct solar power). But nuclear power plants release energy that was stored up from an ancient supernova, the death of some massive anonymous star long before the Sun was born.

It is even possible to estimate how long ago that supernova event took place, or in other words to estimate the age of the heaviest elements in our bodies. This is about 6.6 billion years. The estimate comes from measuring how abundant two particular **isotopes** of uranium are today. Since they decay at different rates, one can work backwards in time to determine how long ago they were equally abundant. The original supernova would have made them in roughly equal quantities, so this gives the estimate of their age. The details of the calculation are in Investigation 11.3 on the following page.

In this section: stars are the factories where elements are made. All elements heavier than hydrogen and helium are made by stars. The process is not straightforward, and would not happen at all if it were not for a strange coincidence of nuclear physics. We owe the evolution of life to the fact that a certain nuclear reaction proceeds much more easily than it should, allowing carbon and oxygen to be formed in stars.

▷ The plural of the word *nova* is *novae* in most languages that use this scientific term. Increasingly in English authors use *novas*. I will remain with the international form in this book.

Investigation 11.3. Finding out how long ago "our" supernova occurred

The supernova that gave birth to the heavier elements in the cloud of gas that eventually condensed to become the Solar System made the heavy elements in our bodies. It is interesting that we can actually work out approximately how long ago that supernova event occurred.

The key is the fact that the supernova made these elements very quickly, within the space of a few minutes. Many of the nuclei that were made then were radioactive. Those that were very unstable decayed into other elements rapidly and disappeared completely. But many radioactive elements decay very slowly, and these are still present today. Two such nuclei are ^{235}U and ^{238}U , two isotopes of uranium. The notation used for identifying nuclei is described in the Introduction.

These two nuclei are not very different, and the processes that formed them were not sensitive to the small effects that make them decay over long time-scales. From the point of view of the physics that dominated the supernova explosion, these two isotopes were stable end-products, produced in essentially the same way. One nucleus got, at random, three more neutrons than the other. Therefore, they were produced in roughly equal numbers. Detailed calculations of the physics of the supernova explosion suggest that the number of ^{235}U nuclei produced was just 1.7 times the number of ^{238}U nuclei. The explosion produced, of course, comparable numbers of ^{236}U and ^{237}U as well, but these are more unstable and have since decayed away.

Since their production, ^{235}U has been decaying slowly with a **half-life** of $7.0 \times 10^8 \text{ y}$, and ^{238}U has been decaying with a half-life of $4.5 \times 10^9 \text{ y}$. The half-life is the time it takes half of the nuclei in a sample to decay. Thus, if there are N_0 nuclei of a certain isotope present at the beginning, then after a time equal to a half-life there are $N_0/2$. After two half-lives, the number is $(N_0/2)/2 = N_0/4$. After k half-lives there are

$$N = N_0/2^k$$

nuclei left. Now, this equation is true even if k is not an integer. So

after, for example, 1.5 half-lives, the number of nuclei has decreased by a factor of $2^{1.5} = 2.83$. The exponent k here is the number of half-lives, or in other words the ratio of the actual time to the half-life. So if we use the letter^a τ to denote the half-life of the isotope and we look at the sample after a time t has elapsed, then we are looking at it after $k = t/\tau$ half-lives. The general equation for the number of nuclei remaining is then our previous equation with k replaced by t/τ :

$$N(t) = N_0/2^{t/\tau}. \quad (11.1)$$

At the present time, the ratio of the number of ^{235}U nuclei to the number of ^{238}U nuclei is about 0.007, as determined from Moon rocks brought back to Earth for analysis by the Apollo astronauts. Let us suppose that we began at time $t = 0$ (the time of the supernova event) with a sample that had 10^{20} nuclei of ^{238}U and 1.7×10^{20} nuclei of ^{235}U . These are in the correct ratio for elements just after the original explosion, but the overall number of 10^{20} is arbitrary. All we will be interested in is ratios of numbers, not the overall number remaining.

We can simply calculate the number remaining and their ratio after successive steps of 10^9 y , using the right value of τ for each element. For ^{235}U the number will be $10^{20}/2^k$, where $k = t/7 \times 10^8 \text{ y}$ is the number of half-life steps in the time t . For ^{238}U it will be the same formula but with $k = t/4.5 \times 10^9 \text{ y}$. The results are summarized in the table below.

What the table shows is that sometime between 6 and 7 billion years after the supernova, the ratio of the two isotopes falls to 0.007, its present value. Therefore, the supernova that gave birth to all our heavy elements occurred between 6 and 7 billion years ago. Since we believe our Sun is only about 5 billion years old, this means that the formation of the Sun did not follow immediately after the supernova: it was apparently not triggered by the supernova. Radioactive dating of other elements shows that the Earth's rocks themselves were formed about 4.5 billion years ago.

	Time after supernova (y)						
	1×10^9	2×10^9	3×10^9	4×10^9	5×10^9	6×10^9	7×10^9
Number of ^{235}U remaining	6.3×10^{19}	2.3×10^{19}	8.7×10^{18}	3.2×10^{18}	1.2×10^{18}	4.5×10^{17}	1.7×10^{17}
Number of ^{238}U remaining	8.6×10^{19}	7.3×10^{19}	6.3×10^{19}	5.4×10^{19}	4.6×10^{19}	4.0×10^{19}	3.4×10^{19}
Ratio $^{235}\text{U}/^{238}\text{U}$	0.74	0.32	0.14	0.059	0.033	0.011	0.0049

Exercise 11.3.1: Getting the age of uranium right

Perform the calculations to fill in the above table, using Equation 11.1 with the two given half-lives. Then take time-steps of 0.1 billion years within the last interval to show that the supernova occurred about 6.6 billion years ago.

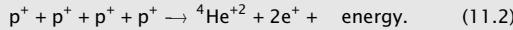
^aThe Greek letter τ is frequently used in mathematical notation to represent particular values of time. It is normally pronounced like "out" backwards, but some speakers say "taw".

If the stars and their explosions are continually making more and more heavy elements, what was the Universe like when the first generation of stars was just forming? By looking for old stars, stars of the first generation that were small enough that their nuclear reactions have not yet run their course (why small stars age slowly will be explained in Chapter 12), astronomers have learned that the gas the first stars formed from was not pure hydrogen: about 20% or so by weight was already helium.

This initial helium must have been made from hydrogen in the early stages of the Big Bang. The bang was hot, very hot, and the same nuclear reactions that now make helium in stars also made helium then. However, the expansion of the Universe cooled the reacting gas off very quickly, freezing out a certain concentration of helium, but quenching the production of heavier elements. Most of the elements of which we are made were left to be cooked up in early generations of stars.

Investigation 11.4. How to make helium

The key nuclear reaction runs something like (but not exactly, as we will see later) the following:



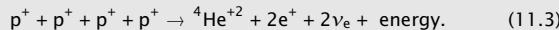
In this equation we use some of the conventional notation of nuclear physics: p^+ denotes a proton, e^+ a positron (a positively charged electron, also called an anti-electron), and ${}^4\text{He}^{+2}$ a nucleus of helium-4. (Helium-4 is the most common form of helium, having a nucleus consisting of two protons and two neutrons.) As is indicated in the equation, the reaction gives off energy, which comes out as the kinetic energy of the products.

A word on the notation in equations like Equation 11.2: we write the name of a nucleus by giving not only the symbol of its element, such as Fe for iron, but also a preceding number to indicate the total number of neutrons and protons in the nucleus. The most common form of iron has 28 protons and 28 neutrons, and so is called ${}^{56}\text{Fe}$. When it is important to indicate the electric charges involved, which for ${}^{56}\text{Fe}$ is 28, we write it as a following number: ${}^{56}\text{Fe}^{+28}$. The charge on the nucleus determines the element it belongs to, so that explicitly showing the charge is redundant; we only do it when necessary for clarity. On the other hand, the number of neutrons in the nucleus is not important for the chemical reactions of the element; two nuclei with the same number of protons but different numbers of neutrons are called *isotopes* of the same element. The preceding number attached to the symbol for the element allows one to distinguish one isotope from another. Some nuclei have alternative names and symbols: the nucleus of ordinary hydrogen is a single proton, also called p or p^+ .

The main thing to watch for in equations like Equation 11.2 is that certain quantities must balance on both sides. For example, the total electric charge of the particle going into the reaction must equal that of the particles coming out, since charge can be neither created nor destroyed. In the above equation, the four protons on the left carry four positive charges. On the right, the helium nucleus contains two charges and the two positrons (called e^+) each carry one, balancing the charge. A similar *conservation law* applies to the total number of protons and neutrons. Protons and neutrons are called **baryons**, and we say that the total baryonic "charge" must balance. The helium nucleus has two protons and two neutrons, each of which has a positive baryonic charge, so the total of protons and neutrons balances as well. This, we shall see, is significant.

Evidently, the nuclear reaction listed above proceeds by changing two protons into neutrons. The balance of electric charge is taken care of by creating two positrons. However, there is another balance law that must be obeyed: **lepton** conservation. Electrons and positrons belong to a class of particles called leptons, and in any reaction the total **lepton** number must not change. Now, electrons have leptonic charge +1, while positrons have leptonic charge -1. The **neutrino** ν is also a lepton, and has leptonic charge +1. Therefore, the creation of two positrons (leptonic charge -2) can be balanced if two neutrinos (leptonic charge +2) are created.

Physicists have discovered that there are in fact *three* families of leptons, each with its own separate balance law. We have been discussing reactions involving the electron family here, so the neutrinos that are created are called **electron neutrinos**, denoted by ν_e . We can now rewrite the reaction in Equation 11.2 in its correct form:



Although it was clear in the 1930s that in principle the Sun could shine this way, the details were not worked out until the 1950s, and

they may still be in need of refinement, as we shall see below. The basic problem is that it is exceedingly rare for four protons to collide all at once, so that the direct reaction given in Equation 11.3 hardly ever happens. Instead, there is a rather complex path through many intermediate reactions, whose end effect is the same as Equation 11.3.

To follow this path, we need to introduce a few more nuclear particles into our discussion. These are **deuterium** – the heavy form of hydrogen that is used to make heavy water – and ${}^3\text{He}$, the light isotope of helium. Deuterium (${}^2\text{H}$) is made from one proton and one neutron. Add one more proton and one gets ${}^3\text{He}$.

There are many ways to formulate a more realistic path to the formation of ${}^4\text{He}$. Pairs of protons collide to form deuterium, emitting a positron and a neutrino. Two deuterium nuclei may then collide to form ${}^4\text{He}$. Or a proton may collide with deuterium to form ${}^3\text{He}$. A fourth proton hits this and forms ${}^4\text{He}$, with the emission of a further positron and neutrino. This network of reactions is called the *p-p chain*.

There are other ways to do the same thing, the most important of which is the so-called carbon cycle, in which nuclear reactions involving carbon nuclei have the net effect of facilitating Equation 11.3, without changing the total number of carbon nuclei; carbon acts as a **catalyst**. The carbon cycle dominates the energy output of stars somewhat more massive than our Sun.

The p-p chain has some side-chains, as well. Nuclei collide and transmute at random, so that some reactions occur in the Sun involving lithium, boron, and other light elements. These are not significant in terms of the energy they contribute to the Sun, but they have assumed great significance in what has become known as the solar neutrino problem. We shall look at this later in this chapter.

The various conserved charges that we have met here all have deep connections with the fundamental forces of physics. Electric charge is responsible for the electromagnetic interactions of particles. Likewise, baryonic charge is associated with the so-called **strong interaction**, which is the glue that binds protons and neutrons together in nuclei, overwhelming the electric repulsion of the protons for one another. The leptonic charges are associated with the weak interaction, which is responsible for beta decay.

There is a strong feeling among modern physicists that in fact all these forces are aspects of a single force, described by a **grand unified theory** of fundamental physics. We already know that the weak interaction and electromagnetic interaction are aspects of the so-called **electroweak** force, and there is strong evidence for unification of this with the strong force. If the forces are not completely separate, then there is the possibility that the conservation laws will not be completely separate, either, so that it may occasionally happen that a single proton will decay into a combination of leptons. Experimental evidence against this happening puts a lower bound on the half-life of the proton of 10^{33} y, so the violation of the conservation law for baryons will be a rare event indeed! Although these considerations do not matter much to the nuclear physics inside stars, they will matter deeply when we come to the most profound questions raised by the modern study of cosmology in Chapter 27.

The discovery that nuclear reactions power the Sun opened up vast new territory in astrophysics. Why should the transmutations stop at helium? Why not carry on and make carbon, oxygen, iron, indeed all the heavier elements? This is indeed what has happened: all the principal elements in our bodies were manufactured in stars that lived and died before the Sun and Earth were formed.

The solar neutrino problem

The main thrust of this chapter is to explore how stars made the elements of which we are composed, and thereby made life possible. We are a side-effect of the nuclear physics that is going on inside stars. The main evidence supporting the picture of nuclear energy generation that we have developed so far is indirect: theoretical calculations based on nuclear physics experiments on the Earth predict models for stars that agree well with their observed properties, such as their luminosities and

In this section: fewer neutrinos from the Sun's nuclear reactions are detected than expected. The reason is becoming clear: neutrinos transform themselves as they move through space between the Sun and the Earth.

temperatures. But direct evidence about the nuclear furnaces in stars is hard to obtain, because we can't see directly into them. The one exception is our Sun. Because the Sun is so near to us, we have the opportunity to look for the neutrinos that the nuclear reactions emit. Elusive as they are, they can be detected. The good news is that they *have* been detected. The bad news is that there aren't as many of them as physicists had expected. This is called the solar neutrino problem.

Earlier in this chapter we met the neutrino, and we learned more about it in Investigation 11.4. A key property of neutrinos is that they have a characteristic called the **lepton number**. This is analogous to electric charge, in that it is *conserved*: the total lepton number of all the particles in a nuclear reaction is the same before and after the reaction.

This leptonic charge is responsible for the nuclear force that physicists call the **weak interaction**, or weak force. When particles carrying a non-zero lepton number collide with other such particles, they can scatter because of the weak force between them. The amount of scattering is much smaller than that which would occur if the particles also had electric charge, which is why the force is called "weak". Neutrinos do not have electric charge, so they can scatter from other particles *only* via the weak force. The result is that neutrinos produced in the center of the Sun stand little chance of ever scattering off anything else on their way out. Clearly, they will pass right through the Earth almost unhindered as well. This makes them excellent probes of the central conditions in the Sun: when we detect solar neutrinos on the Earth, we "see" directly into the core of the Sun itself.

The problem with detecting solar neutrinos is the other side of the same coin: they don't interact much with matter. The flux of neutrinos itself is huge. Every second something like ten billion solar neutrinos (10^{10}) pass through your body. But they just pass through, leaving almost no energy behind, inducing almost no nuclear reactions. So they will also pass through any detector we might build almost without noticing it. Physicists must use extremely sensitive techniques to find solar neutrinos at all.

►Contrast the easy ride that neutrinos have with the plight of photons in the Sun. We saw in Chapter 8 that a photon scatters more than 10^{20} times on its way out from the center! This illustrates the feebleness of the weak interaction compared with the electric forces.

►Cosmic rays are fast-moving protons that fly around the Milky Way. Most seem to be particles ejected from supernova explosions long ago, which follow random paths in and around the Milky Way, guided by our Galaxy's weak magnetic field. Some cosmic rays have energies above 10^{20} eV, which makes them the most energetic particles scientists have ever dealt with. The origin of these particles is a deep mystery, which we will explore in Chapter 27.

The first solar neutrino observations were made by the American physicist Raymond Davis (b. 1914) in the Homestake Gold Mine in South Dakota. He went below ground into the mine in order to use the rocks above him to screen out **cosmic rays** (high-speed particles hitting the Earth from space), which might otherwise have obscured the signal from the neutrinos. The basis of the experiment is a reaction in which a solar neutrino transforms a chlorine nucleus into one of a radioactive isotope of argon. The chlorine is in liquid form inside the experimental chamber, while the argon is a gas. By extracting the gas periodically and counting the number of radioactive argon decays, and by using the known rates at which neutrinos will interact with chlorine, Davis is able to infer the neutrino flux falling on the detector and hence the neutrino luminosity of the Sun. Davis has created a new unit for measuring the flux of neutrinos, the **Solar Neutrino Unit**, abbreviated snu. One snu represents one captured neutrino for every 10^{36} target chlorine atom in each second.

This is an experiment of extraordinary delicacy, but over more than 20 years of collecting data, Davis has consistently measured a capture rate of about 2.5 snu, which is a factor of two or three below the predictions of the Standard Model of the Sun's interior. The Standard Model starts from considerations like those in Investigation 8.5 on page 94, where we built a model of the Sun. To this rather simple approach, physicists add the best information available about the nuclear reactions that can take place, and carefully take account of the radiation pressure provided by the outgoing radiation. When they build a model that has the mass,

luminosity, size, and composition of the Sun, it predicts a neutrino luminosity about three times larger than Davis sees.

Such a discrepancy is very serious, and indicates that there is something we don't understand about either the Sun or neutrinos. The interpretation of Davis' experiment is made more complex by the fact that he does not detect neutrinos from the p-p chain, which supplies almost all the energy radiated by the Sun, but from a minor side reaction. This is because the nuclear reaction Davis uses in his detector requires the incoming neutrino to have a relatively large energy, and the p-p neutrinos are not energetic enough. The only neutrinos produced in the Sun that have enough energy are produced as a result of one of the inconsequential side reactions that are going on all the time. The most important one is:



In turn, the boron nucleus ${}^8\text{B}$ is unstable to a form of **beta decay**, emitting a positron and a particularly energetic neutrino when it decays back to ${}^8\text{Be}$. This reaction does not contribute significantly to the energy output of the Sun, but the rate at which it proceeds, and hence the flux of neutrinos it produces at the Earth, depends sensitively on the local conditions in the solar core, especially the temperature and density.

For some time, Davis' experiment was the only one able to detect solar neutrinos, so there was always a possibility that there was some flaw in the experiment or in the nuclear physics calculations that are required to interpret it. Then an independent measurement of the neutrino flux was made by the Kamiokande neutrino detector in Japan. This detector was built to look for possible spontaneous decays of the proton, which we refer to in Investigation 11.4 on page 127. The detector entered the realm of astrophysics when it registered neutrinos from the supernova explosion called **SN1987A** that occurred in the Large Magellanic Cloud in February 1987. I shall have more to say about the neutrinos from **SN1987A** and the Kamiokande detector in Chapter 12. For now, the important thing is that this detector is also sensitive only to high-energy neutrinos, so it directly tests Davis' experiment. Within the experimental errors, it has confirmed the shortfall in the neutrino flux.

What could be the cause of this deficit? Physicists first tried to modify the Standard Model of the Sun, reasoning that we know less about the solar interior than about the nuclear physics that can be tested in the laboratory. But they have found that changing the flux expected from the Sun by such a large amount is hard to do. In particular, the evidence from helioseismology (Chapter 8) places very strong constraints on the temperature and density profiles inside the Sun, and made it unlikely that the solution to the problem could be purely astrophysical.

A second possibility is that the rates of some of the minor nuclear reactions inside the Sun have been overestimated because we don't understand the nuclear physics itself as accurately as we think. Nuclear experiments are underway in order to test this understanding. There have recently been revisions in some reaction rates, and these have helped reduce the gap between theory and observation. But a substantial difference still remains.

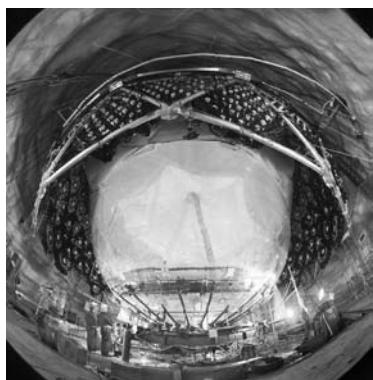
However, there is a third, very interesting, class of solutions that has been proposed, and which has very recently gained strong observational support. The idea is that something happens to the neutrinos during their flight from the Sun to the Earth. At first one might try out the idea that they could simply decay into something else. Then the Sun might be producing the right number of them, but they don't reach us. Unfortunately, this is not consistent with the fact that Kamiokande saw just about as many neutrinos from the supernova in 1987 as scientists expected,

assuming they do not decay. If half of them might decay just going from the Sun to the Earth, then essentially all of them would have disappeared coming from the supernova, which occurred 10^{10} times further away! So we have learned from the supernova that neutrinos do not simply disappear even as they travel over the vast distances between stars.

A more subtle variation on the decay idea, however, does seem to be happening: neutrinos change into other neutrinos and back again, in an oscillating manner, as they move through space. In Investigation 11.4 on page 127 we pointed out that there are actually *three* types of neutrino, each associated with one of the three charged leptons. However, only the electron neutrino will trigger the nuclear reactions in Davis' detector. If electron neutrinos transmute into other varieties between the center of the Sun and here, then this would explain the low observed flux. If they oscillate back again over a longer journey, and if each type of neutrino does this at a different rate, then there would be no contradiction with the neutrino observations of SN1987A. All it would mean is that the observed neutrinos were about one third of the original number, and since scientists have only a very approximate idea of the number of neutrinos to expect from a supernova, this would be an acceptable solution.

The neutrino oscillation solution violates the separate laws of lepton number conservation, but we would have to accept this; at least the *total* lepton number would still be conserved. The issue is one for experimental physics, and theory will have to follow. A number of detectors have recently been built to look for solar neutrinos with different energies, from different parts of the nuclear reaction chain. Many of them use gallium as a target rather than chlorine. They have measured capture rates about half of the prediction of the Standard Solar Model. The Kamiokande detector has been replaced by a much larger one, called SuperKamiokande. It has measured a solar neutrino flux of $2.32 \times 10^6 \text{ cm}^{-2} \text{ s}^{-1}$, again about half of the prediction of the Standard Solar Model. These results are all consistent with the Davis measurement, since they measure neutrinos from different reactions. They all reinforce the view that neutrino physics is responsible for the puzzle.

Figure 11.2. A fish-eye-lens view of the SNO detector. The inner white vessel holds 1000 tonnes of heavy water (in which deuterium replaces normal hydrogen), with which the neutrinos interact. The metal cage is the top part of the structure holding phototubes that register light emitted by particles produced by the neutrinos. Note the scale of the people in the photo. (Photo courtesy Ernest Orlando Lawrence Berkeley National Laboratory.)



sensitive to some of the non-electron neutrinos, and since the Sun should be producing only electron neutrinos, this implies that some of the produced neutrinos have changed into other types by the time they get to the Earth. Physicists have estimated from the details of the two experiments that the original flux of electron neutrinos leaving the Sun should be $5.44 \times 10^6 \text{ cm}^{-2} \text{ s}^{-1}$, in excellent agreement with the predictions of the Standard Solar Model.

At the time this book is being written (2002), scientists have just had first results from a new heavy-water detector. Called the Sudbury Neutrino Observatory (SNO), it is located in a mine owned by the Inco Mining Company in Sudbury, Ontario. This detector has the ability to distinguish between different types of neutrinos, and it should finally point the way to a solution. The first data from SNO are a measurement of the electron neutrino flux from the Sun. The flux is less than SuperKamiokande, only $1.75 \times 10^6 \text{ cm}^{-2} \text{ s}^{-1}$. Since SuperKamiokande is

Further observations expected from SNO, including its own direct measurements of the total neutrino flux, should settle the matter and point the way to measuring values for the different oscillation wavelengths. The implications for theories of fundamental physics are only just beginning to be assessed. One implication is already clear, however: the neutrino must have a small mass. It cannot be a massless particle. This is another consequence of special relativity. We will see in Chapter 15 that massless particles travel at the speed of light, and that particles moving at the speed of light experience no lapse of time: time stands still for them, and if they had an internal clock it would not advance at all. No dynamical process, like oscillation from one type of neutrino into another, could happen; nothing at all could change for a massless neutrino. Other experiments have already shown strong evidence for tiny neutrino masses, less than one millionth of the mass of an electron. Neutrino oscillations also require masses of this order.

Life came from the stars, but would you have bet on it?

Viewed from different perspectives, the evolution of life on Earth can seem either almost inevitable or wildly improbable. Stellar astronomy forms the background against which the story of life unfolds. Consider the astronomical ingredients. Stars had to form before the Sun, make elements heavier than helium, and return a good fraction of them back to the interstellar clouds of gas from which the Sun's generation of stars would form. The Sun had to condense from the cloud of gas, leaving a sufficiently massive disk of gas around it from which the planets formed. The Sun needed to be hot enough to warm the planets, but not so hot that it would exhaust its nuclear fuel before life had time to evolve.

None of these ingredients is very unusual. Star formation will require many generations to exhaust the hydrogen supply, so there will be a long time in which stars are forming from clouds seeded with heavy elements. There is increasing evidence that many, perhaps most, stars like the Sun have left disks behind, from which planets might form. Indeed, surveys using specially designed telescopes are now turning up many planets around nearby stars. And the Sun is a very ordinary star: there are many like it that will shine steadily for the billions of years apparently required for life to evolve. These minimal conditions for life probably exist in billions of places just in our own Milky Way galaxy.

Life as we know it also probably required that the Earth have just the right distance from the Sun, and just the right mass and composition to allow things like volcanism and plate tectonics (continental drift) to continue over billions of years. This is somewhat more special, but it still might not be very surprising if there were millions of places in the Milky Way where life could evolve, and possibly is doing so right now. It is certainly not possible to estimate accurately the likelihood of this. The most famous attempt to do so, by Frank Drake (b. 1930), resulted in an equation full of undetermined factors (see Figure 11.3 on the next page).

Yet there is much that seems less probable if we look further back in time. When we end our study of cosmology in Chapter 27, we will see that we have no real explanation for why the Universe is the age it is. It might have happened that the Big Bang was not quite so big; then the expanding material of the early Universe might have turned around and re-contracted after only, say, a few million years. This would not have allowed enough time for life to evolve anywhere. We will also see, more dramatically, that the very laws of physics as we know them today took shape in the very early Universe. The mass of the proton and electron, the exact relative strengths of the nuclear and electromagnetic forces, the sizes of the density irregularities that eventually led to the formation of stars and galaxies in the expanding Universe: all of these things may have been determined essentially

>As this book was being finished, the 2002 Nobel Prize for Physics was announced. It was shared by Davis and Masatoshi Koshiba (b. 1926), founder of the Kamiokande experiment, with Riccardo Giacconi (b. 1931), a pioneer of X-ray astronomy.

In this section: the improbability of life is a subject of intense interest to many people. The stars provided the raw materials, the Sun provided the nursery. But the laws of physics seem to have a lot of fine-tuning that allowed just the right conditions to be present for life as we understand it.

Figure 11.3. The Drake equation for estimating the probability that life could have evolved in our Galaxy could also be applied to equally difficult problems!
Reprinted with kind permission of Mark Heath.



Using Frank Drake's famous equation, Betty calculates the probability of finding intelligent life on a Saturday night.

at random just after the Big Bang.

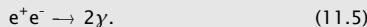
This is relevant, because the evolution of life seems to depend on a little "fine-tuning" of the fundamental physics here and there. For example, we noted above that heavier elements form in stars when helium nuclei combine to form carbon, oxygen, and so on. There is a deep mystery, an incredible coincidence, underneath that bland statement. Carbon has 12 particles in its nucleus: six protons and six neutrons. It therefore requires three helium nuclei as building blocks. Now, if one relies on random collisions to drive nuclear reactions, it is very improbable that three alpha particles will converge at the same place at the same time. Ordinarily, one would expect that two would collide, forming a nucleus with eight particles in it, and then some time later a third helium nucleus would collide with the eight-particle nucleus to form carbon. Unfortunately, *there are no stable nuclei with eight particles*. Any such nucleus formed from two alpha particles will immediately disintegrate. In the conditions inside stars, such objects do not last long enough for a third alpha particle to come by. Nor was it any more likely in the Big Bang, which we will study in Chapter 25. When physicists first began to study these things, it seemed there was no way to explain elements heavier than helium.

This bottleneck would have prevented the formation of planets and the evolution of life, if it were not for an apparent accident. In a brilliant flash of insight, Hoyle realized that there was another possibility. *If* three alpha particles had a bit more attraction for one another than one would expect on general grounds, then the rate of three-particle reactions would be higher: the particles would not have to come quite so close to one another to get drawn into the reaction that forms carbon. He calculated that if there was a sufficiently strong extra attraction, then it would show up as what physicists call a long-lived **excited state** of the carbon nucleus. This means that if one were to fire another particle (maybe another alpha particle) with

Investigation 11.5. Where do the photons come from?

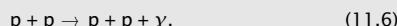
The Sun radiates light to us, yet in the nuclear reactions of Investigation 11.4 on page 127 we have not seen any that produce photons (gamma-rays). Where does the light come from?

There are two aspects to the question: how the photons are generated deep inside the Sun, and then how they get to the surface. In the core of the Sun, there are two places where photons can be made. The first is as a result of the production of positrons, the e^+ particles in, for example, Equation 11.3 on page 127. The positrons quickly meet electrons, which are their anti-particles, and annihilate via the reaction



Each of the two photons is very energetic, having about 0.5 MeV of energy. (The symbol MeV denotes "million electron volts", which is 10^6 eV = 1.6×10^{-13} J.) These photons quickly lose their energy, however, in the Compton scattering process we shall describe below.

The second way that photons are generated in the central core is simply by the collision of charged particles. For example, most of the time that two protons collide they do not produce deuterium; instead, they simply deflect each other's motion and change some of their energy into a photon:



Exercise 11.5.1: Entropy of the Sun

We have seen in Chapter 8 that a typical photon in the Sun takes 10^6 years to randomly "walk" out of the Sun. That means that the Sun contains all the photon energy it generated by nuclear reactions in the last million years. This must all be in the form of photons, since the particles in the Sun have the same total energy today as they had a million years ago. (a) Calculate from the solar luminosity how much energy the Sun contains in photons. (b) If the average temperature inside the Sun is 10^5 K, calculate mean energy of each photon. (c) From these two results estimate the number of photons inside the Sun. (d) From the mass of the Sun, assuming for simplicity that it is composed entirely of hydrogen, calculate the number of protons (hydrogen nuclei) in the Sun. (e) Find the ratio of the number of photons to the number of protons in the Sun. This is a measure of what physicists call the **entropy** of the Sun.

just the right energy at a carbon nucleus, the impact would cause it to split up into three alphas, but these would not fly apart; the extra attraction would hold them together, oscillating about one another, until they eventually emitted a gamma-ray photon and settled back down to a normal carbon nucleus.

Experiments soon found the predicted effect. Carbon *does* have such an excited state, three alpha particles *do* attract one another more than one might at first expect, and carbon *can* form inside stars by three-alpha collisions. The synthesis of elements beyond carbon then needs only a succession of two-particle reactions: add one helium nucleus to carbon and one gets stable oxygen; combine oxygen and carbon nuclei and one gets stable silicon, and so on. So life seems to hinge on the existence of this one excited state of carbon. This seems to be a small detail of the laws of physics. If certain fundamental numbers, like the mass of the electron, the unit of electric charge, or the strength of the nuclear force were to have been just slightly different, then life would simply not have been possible: the elemental building blocks would not have been there.

Another example of the special nature of the laws of physics is given in the next chapter, where we consider the death of stars by supernova explosions. Some of these are triggered by the collapse of the core of a massive star when it runs out of nuclear fuel. The collapse releases energy, and this blows the rest of the star apart. These explosions are one of the ways that elements heavier than helium are placed into clouds of interstellar gas, ready to act as seeds for planets and for life in the next generation of stars. It appears that our own Solar System formed from gas enriched by such an explosion, without which we would not be here.

It seems, from theoretical calculations, that it is not easy to blow such stars apart. The energy released by such a collapse is carried away by neutrinos, which leave only just barely enough energy to blow the envelope of the star away. Now, we

This slowing down of the protons gives rise to the physicists' name for this reaction, **bremsstrahlung**, from the German *bremsen* (to brake) and *Strahlung* (radiation).

In addition, photons are constantly running into protons and scattering off them, exchanging a little energy each time. This is called **Compton scattering**. It can, of course, occur between photons and electrons or other charged particles as well. Compton scattering and bremsstrahlung keep the photon gas at the same temperature as the particles, i.e. they insure that the typical energy of a photon equals the typical energy of a proton or electron. In particular, the gamma-ray photons produced by positron annihilation quickly lose their excess energy this way. They dissolve into the general background gas of photons.

In this manner, we say that the interior of the Sun generates a photon gas that is in thermal equilibrium with the particles. The overall temperature of this equilibrium is determined by gravity, by the pressure required to hold up the whole mass of the star against its gravitational self-attraction.

shall see that this collapse only occurs when the mass of the exhausted core reaches about one solar mass. This mass can be calculated, quite remarkably, from simple fundamental constants of nature: Newton's constant of gravitation G , Planck's constant h , the speed of light c , and the mass of the proton m_p . If the proton were a bit more massive, then (as we shall see) the core would collapse when it had less than one solar mass. Such a collapse would release less energy, and perhaps not enough would be available to blow apart the star. What is more, the collapse causes an explosion only because something halts the collapse, causing a rebound of the infalling material. This something is the formation of a neutron star, which we will also study in Chapter 12. We will see there that the existence of neutron stars depends on the exact strength of the nuclear forces. A small weakening of these forces would have led the collapse to form black holes, with little or no rebound, and the interstellar medium would not have been enriched as much as it has been.

This sort of fine-tuning can be found elsewhere, as well, and we will give more examples in the final chapter. This has given rise to a point of view about the history of the Universe that is called the **Anthropic Principle**. The mildest form of this principle holds that, since we are part of life, any universe in which the fine-tuning prevented life from forming would not have had us in it to puzzle over the fine-tuning. Therefore, the fine-tuning is no puzzle; it must be taken for granted from the simple fact that we are here to discover it. This point of view seems plausible if one believes that there might be many "universes", so that the cosmological experiments can be repeated many times, and there is nothing special about ones that happen to produce life on obscure planets. One might imagine a repeating Big Bang, endlessly cycling through expansion and re-collapse (the "Big Crunch"). Or one might imagine the Universe to be extremely large, and that different regions of it have different values of such things as the proton mass. We would then not see such regions because light has not had time to reach us from them since the Big Bang.

A more radical version of the Anthropic Principle is more metaphysical: the Universe is fine-tuned to produce life because its *purpose* is to produce life. Scientifically, this could be a dead end, since it discourages further questions about the fine-tuning, except possibly to try to find relations between examples of it that seem rather distant from one another. I prefer the first version of the principle, especially since, as we will see in Chapter 27, many physicists are working hard now to arrive at a theory of quantum gravity, and this may well predict that our Universe is not the only one, that the values of the fundamental constants are not the only ones that have occurred or will occur.

I have not addressed here the other ingredients necessary for life, especially the enormous chain of chemical reactions building upon one another to make the complex molecules of life on the present Earth. It is usually assumed that this happened spontaneously on the Earth, using a combination of catalysts and natural selection to guide the chemistry. Some astronomers, including Hoyle, have revived the idea of **panspermia**, which is that life could have spread from one star to another, perhaps propelled by strong mass flows that astronomers observe from some kinds of giant stars, so that it need not have arisen on Earth. That still requires that the complex chemistry take place somewhere in the Galaxy, but it allows biologists to look in environments very different from the early Earth for the very first steps toward evolution.

Birth to death: the life cycle of the stars

The cycle of birth, aging, death, and re-birth of stars dominates the activity of ordinary galaxies like our own Milky Way. The cycle generates the elements of which our own bodies are made, produces spectacular explosions called supernovae, and leaves behind “cinders”: remnants of stars that will usually no longer participate in the cycle. We call these white dwarfs, neutron stars, and black holes.

Governing this cycle is, as everywhere, gravity. An imbalance between gravity and heat in a transparent gas cloud leads to star formation. The long stable life of a star is a robust balance between nuclear energy generation and gravity. This balance is finally lost when the star runs out of nuclear fuel, leading to a quiet death as a white dwarf or to a violent death as a supernova.

Even the cinders, all unusual objects, can be understood from simple calculations based on elementary physical ideas. We have already met black holes. White dwarfs and neutron stars exist in a balance between gravity and quantum effects: they illustrate the deepest principles of quantum theory. Exotic as these may seem, life on Earth would not exist without the neutron stars and white dwarfs of our Galaxy.

Starbirth

The Milky Way is filled not only with stars but with giant clouds of gas that have not yet formed stars. We call these clouds “molecular clouds” because they are dense enough for chemical reactions to take place to form molecules. The chemical elements in the clouds were put there by earlier generations of stars, by processes we will study later in this chapter. Such clouds contain many simple molecules, mostly molecular hydrogen (H_2 , with traces of carbon monoxide and formaldehyde), and they also contain solid grains of interstellar dust.

We met dust grains in Chapter 7; they are the raw material of which the planets were made. Dust also obscures the astronomer’s view of distant stars. There is a small amount of dust in any direction we look, scattering light and making stars seem redder than they really are, just as dust in the Earth’s atmosphere makes the Sun seem red at sunrise and sunset. In some molecular clouds, the dust is so thick that it obscures everything beyond it, sometimes sculpting spectacular shapes in the night sky (see Figure 12.1 on the following page).

The chemistry of clouds is interesting in its own right, and sometimes leads to spectacular consequences. One of the most extraordinary ones is the interstellar **maser**. A maser is like a **laser**, but the emission comes out in radio waves and not in light. Because clouds are rarified, collisions among molecules are rare. More common are collisions with cosmic rays. These act as the “pump” which puts energy into the maser; in the right circumstances, the result can be that clouds continuously emit pencil-thin, intense beams of radio waves in random directions. Scientists developed masers and lasers in the laboratory only relatively recently, but Nature has been producing them for billions of years!

In this chapter: stars form in molecular clouds and die when they burn up their fuel. Small stars die quietly as white dwarfs, larger stars explode as supernovae. In both cases, they return some of their material to the interstellar medium so that new stars and planets can form. White dwarfs, and the neutron stars that usually form in supernova explosions, are remarkable objects. They are supported against gravity by purely quantum effects, so they do not need nuclear reactions or heat to keep their structure. We learn about the quantum principles involved and use them to calculate the size and maximum mass of white dwarfs.

In this section: stars form when portions of gas clouds collapse. The criterion governing collapse is the Jeans criterion, which we derive.

►The figure underlying the text on this page is from a computer simulation of a supernova explosion. The gas, rotating about a horizontal axis in this image, has “bounced” from the neutron star core (center left) and is moving outwards through the envelope with turbulent convection. From a paper by K Kifondis, T Plewa, and E Müller in AIP Conf. Proc. 561: Symposium on Nuclear Physics IV (New York, 2001). Used with permission of the authors.



Figure 12.1. The Eagle Nebula (also known as M16) is a good example of a cloud of gas and dust. The stars visible in the pillars are nearer to us than the dust; everything further away is blocked out. Light from very young stars that have formed here is eroding the pillars. Recent infrared observations have revealed stars forming inside the pillars, although not inside the tiny prominences, as was at first believed. (Photo courtesy NASA/STScI.)

If a cloud is sufficiently dense or sufficiently cold, parts of it can begin to contract to form stars. There is a minimum size for a region that will contract: it has to have enough mass and enough gravity to overcome the pressure in the cloud. This minimum size is called the **Jeans length**, because it was first determined by the British astrophysicist Sir James Jeans (1877–1946). It is given by

$$\lambda_{\text{Jeans}} = \left(\frac{\pi k T}{G \rho m} \right)^{1/2}, \quad (12.1)$$

►Actually, this equation is a special case of the Jeans formula, where we make the assumption that the temperature T of the gas does not change while it contracts. This is the case for clouds that are unable to trap heat until they get much denser.

We work out this size in Investigation 12.1.

Once parts of a dense molecular cloud start to contract, starbirth gets underway. Astronomers call the contracting gas a **protostar**. Astronomers see many clouds where stars are forming. One is illustrated in Figure 12.1. There are probably several things that can provide the initial disturbance that triggers the contraction of a region in a Jeans-unstable cloud: explosions of supernovae of stars already in or near the cloud, collisions between two clouds, or the compression of a cloud as it moves into a region of stronger gravitational field in the Milky Way (i.e. through the plane of our spiral galaxy). Simulations of star formation using supercomputers are beginning to shed light on these mechanisms and on what happens when stars begin to form (Figure 12.2 on page 138). These simulations are, unfortunately, far beyond the scope of our Java programs!

As the protostar contracts, it eventually becomes dense enough to trap radiation and begin to behave like a black body. At this point its temperature rises sharply

Investigation 12.1. Forming stars by the Jeans instability

The key insight into how stars form from molecular clouds is simply to understand that the cloud is *not* a star! This may seem obvious, but in physical terms it means that the cloud is not a black body: it does not trap radiation, but instead is transparent to photons. In Chapter 8 we discussed the stability of a star, and we found that the pressure in a star could balance gravity when the star is compressed if the polytropic index was larger than 4/3. Here we study the same balancing act, only in a transparent cloud.

In a long-lived molecular cloud, the balance between gravity and gas pressure has to be maintained without internal energy generation. Thin clouds have low temperatures, typically 20 K, where energy input from cosmic rays coming into the cloud or heating by nearby hot stars can balance the energy losses from emitting photons. These photons are released when molecules of the cloud vibrate, say after encountering a cosmic ray or colliding with another molecule. The molecules emit low-energy photons in the **microwave** or **sub-millimeter** parts of the spectrum. These low-energy photons are not scattered by the dust in the cloud, so the cloud is truly transparent at these wavelengths, despite being opaque to optical light, as in Figure 12.1. In this way a cloud can be kept at a fixed temperature despite any disturbances, unlike what would happen if the cloud trapped its photons like a black-body star.

In such a cloud, the key balance is between the random kinetic energy of the molecules and their gravitational potential energy. If the cloud contains a region of size R and mass M that is large enough that molecules moving with their thermal motion (at the cloud's temperature T) do not have the escape speed from that region, then roughly speaking they are trapped by gravity, and the region will begin to collapse if any small disturbance gives it an inward push. On the other hand, if molecules can escape from the region then a disturbance will not have time to collapse before the molecules diffuse away. To find approximately the size R at which a spherical region becomes unstable to collapse we simple set the average kinetic energy $3kT/2$ of a molecule of mass m equal to (the absolute value of) its gravitational potential energy GMm/R , and then replace the mass M of the region by its expression in terms of its density ρ , $M = 4\pi\rho R^3/3$. This gives

$$\frac{3}{2}kT = \frac{1}{3}\pi G\rho R^2.$$

Exercise 12.1.1: The conditions for star formation

A typical molecular cloud has a temperature $T = 20$ K, a composition mainly of molecules of H₂ (molecular hydrogen), and a density that corresponds to having only 10^9 molecules of H₂ per cubic meter. Calculate the Jeans length and jeans mass of this cloud. Compare the mass you get with the mass of the Sun.

If we solve for R we get

$$R = \left(\frac{9kT}{8\pi G\rho m} \right)^{1/2}.$$

Our argument is a bit crude, since molecules near the surface of the region can leave the region even with smaller speeds, which will diffuse the contraction. A more sophisticated mathematical analysis first performed by Jeans leads to the slightly larger **Jeans length** λ_J :

$$\lambda_J = \left(\frac{\pi kT}{G\rho m} \right)^{1/2}. \quad (12.2)$$

This is about three times larger than our estimate, but it has the same dependence on temperature, density, and molecular mass, so this indicates that our physical analysis is correct.

Any part of the cloud larger than this size has enough self-gravity to collapse. The mass of this region is $4\pi\rho\lambda_J^3/3$, which is called the Jeans mass

$$M_J = \frac{4\pi}{3}\rho\lambda_J^3 = \frac{4\pi^{5/2}}{3} \left(\frac{kT}{Gm} \right)^{3/2} \rho^{-1/2}. \quad (12.3)$$

The important part of this formula is that M_J decreases as ρ increases. Thus, as the unstable region of initial size λ_J collapses, its density rises. This lowers the Jeans mass and makes smaller parts of the cloud unstable to collapse. If, realistically, the density is not uniform in the cloud, then a collapsing region is likely to *fragment* into many smaller regions, and this fragmentation could occur on many scales. It only stops happening when the cloud cannot maintain its original temperature, either because it becomes less transparent or because it is heated by the contraction faster than the molecules can radiate energy away. At this point the temperature can rise, and the conditions for further instability and fragmentation become more like those we discussed for stars in Chapter 8.

and it begins to shine with visible light. The energy released by contraction, which we calculated in Investigation 11.1 on page 123, continues to be radiated away so the star can continue to contract. Although this energy is not sufficient to power the Sun for billions of years, the protostar can shine for a million years or so on it. (See Exercise 11.1.2 on page 123 for the calculation.)

The gravitational thermostat

Once a star is formed, it will live in a fairly steady fashion for a very long time. We can't understand the death of the star until we understand how it manages to live quietly for as long as it does.

The nuclear reactions that take place in the Sun also take place in thermonuclear explosions, commonly called hydrogen bombs. Why is it then that the Sun has not blown itself up: how can it burn hydrogen in such a steady way? The answer is in the way gravity holds the black-body star together.

Suppose something happened inside a star to make the reactions proceed faster, such as a small increase in the temperature. Then the extra energy released (as photons and in fast particles) would immediately tend to expand the central part of the star, reducing the temperature and density, and thereby reducing the rate of

In this section: the luminosity of a star remains steady, despite all the turbulence inside, because gravity and gas pressure strike a cooperative balance.



Figure 12.2. Four snapshots from a computer simulation of the formation of stars as a result of the collision of two clouds of gas. The stars form along filaments, and often form binary and triple pairs.

Images courtesy A P Whitworth,
Cardiff University.

In this section: most of the lifetime of a star is spent on the main sequence, which means that it resembles the Sun.

reactions. If the opposite happened, say a temperature decrease, then the reactions would put out less energy, there would be less pressure in the gas, gravity would make the star contract and heat up, and the reactions would go faster again.

The self-corrections provided by gravity keep the nuclear reactions in a star in a steady state.

This kind of self-correction does not happen in a molecular cloud, which is transparent to radiation, and that is why the Jeans instability grows. Eventually it leads to the formation of stars that *are able* to maintain this balance.

This kind of self-correction is also much harder if one does not have self-gravity. In the thermonuclear bomb, there is no attempt to sustain the reactions. Instead, the material is arranged so that nearly all the desired reactions take place before the released energy has a chance to scatter the material. In a nuclear fusion reactor, the only way to keep the material together for long enough is by using magnetic fields. This is very difficult to do, and experimenters have found that there are many situations in which a small change in the density or reaction rate does not get corrected naturally.

The goal of fusion research is to find a way to sustain the reactions at a high enough temperature and density to get more energy out than has been put in to make the magnetic field and heat the gas. This goal is getting nearer, but progress is slow, principally because fusion reactors do not have the natural thermostat provided by self-gravity.

The main sequence

We have seen in the previous chapter that stars “burn” hydrogen to make helium. The rate at which they use the hydrogen, and hence the luminosity of the star, depends on how massive the star is. More massive stars are hotter in the center, which means that nuclear reactions will proceed faster, and the star will shine more brightly. Such stars also have more hydrogen to burn, of course, before they end their normal life, but it turns out that their luminosity increases so strongly with their mass that their lifetime actually shortens.

The more massive the star, the shorter will be its life before it reaches the end of its hydrogen-burning time.

In Figure 12.3 I have plotted the result of theoretical calculations of both the luminosity and lifetimes of normal stars. Notice that the lifetime of massive stars is as short as 10 million years, only 1/1000th of the Sun’s lifetime. This figure shows another interesting curve: the amount of energy released by an average kilogram of mass of each star over the life of the star is roughly the same for all stars.

Stars differ in the rate at which they extract energy, but all of them get about the same amount out of each kilogram in the end. This shows us that stars live as ordinary stars until they have exhausted their nuclear fuel. It is the end of nuclear burning and not some other cause that brings about the death of a star.

In Investigation 12.2 on page 141 we extract some numerical information from this figure.

Although stars spend long periods of time in a fairly steady state, burning with the luminosity shown in Figure 12.3, they do evolve slowly as their interior composition changes and as they lose mass. The Sun is losing mass at a very slow rate at the moment, but it may have lost a significant fraction of its original mass in the first billion years after it formed, and it is likely to lose almost half of its mass as it evolves into a red giant, some 5 billion years from now. Composition changes in the Sun may also affect its luminosity, which is probably very slowly increasing.

We show in Investigation 12.2 on page 141 that the luminosity L of a star is a very sensitive function of its surface temperature T : $L \propto T^8$. This can be seen in direct observations. Recall that astronomers can measure the surface temperature of a star by measuring its color. They cannot directly measure its absolute luminosity because the distance to a star is usually unknown, but if we consider only stars that are in a given cluster of stars, all at the same distance from us, then the ratios of their absolute luminosities will be the same as that of their apparent brightnesses. Since the luminosity–temperature relation is only a proportionality, it should then be true that for stars in a given **star cluster**, (apparent brightness) \propto (surface temperature)⁸.

In Figure 12.4 on the next page we display such a plot for the stars of a single cluster. Astronomers call this plot the **Hertzsprung–Russell diagram**, after the two astronomers who independently devised it, the Dane Ejnar Hertzsprung (1873–1967) and the American Henry Norris Russell (1877–1957). The normal stars that we have been describing lie in a diagonal band with a slope that is consistent with the proportionality we have deduced. This band of stars is called the *main sequence*. Notice, however, that there are well-defined bands off the main sequence that also contain a large number of stars. These must be stars which are not in the normal stage of their life cycle.

Some of these are newly-formed stars (pre-main-sequence stars) that are shining by gravitational contraction, but most are stars which have finished hydrogen burning and have moved off the main sequence. The behavior of stars as they leave the main sequence is very complex, and even now is not fully understood, although it can be modeled quite well on a computer. (This sort of computer program would also be beyond our scope here!) But again gravity plays a crucial regulatory role.

Giants

Consider the stars above and to the right of the main sequence. They have generally lower temperatures yet larger luminosities than main sequence stars. This means they must be much larger in size: they are called *giants*. Here is how stars become giants.

When a star exhausts the hydrogen in its center, there is little energy being produced to hold up the weight of the outer parts of the star (which contain, after all, 90% or more of its mass). The inner part of the star contracts and heats up. What happens next depends sensitively on how much mass the star has and what its exact composition is: it is rather sensitive to how much carbon, oxygen, and so on were in the initial cloud that contracted to form the star. Various things can happen in various combinations, as follows.

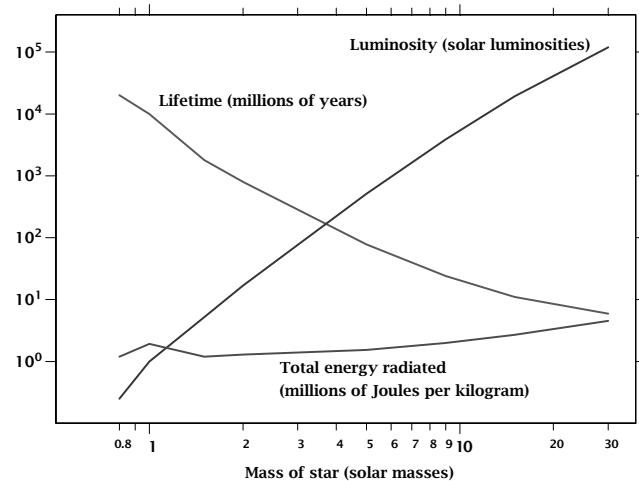


Figure 12.3. Luminosity and lifetime of normal stars. Stars of larger mass have higher luminosities and shorter lives. The energy released by an average kilogram of the star is, however, essentially the same for all stars. All three curves are referred to the scale on the left, using the units indicated for each curve.

In this section: when stars run out of hydrogen to burn, they change their interior structure, and become giants. We examine the causes.

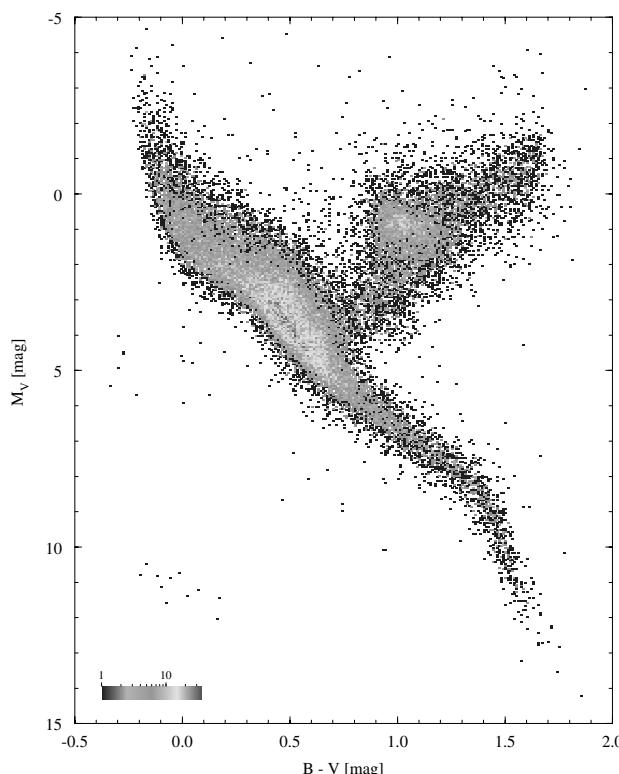


Figure 12.4. The temperature-luminosity diagram of stars measured by the Hipparcos satellite, mentioned in Chapter 9. The vertical axis is the absolute visual magnitude, which measures the luminosity of the star. The horizontal axis is the color of the star, measured by $B - V$, the difference between the blue and visual magnitudes. The Hipparcos survey parallaxes gave distances to stars accurately enough to deduce good values for their absolute magnitudes. Image courtesy esa.

▷ Astronomers use the word **nebula** for any diffuse cloud of gas around a star.

In this section: once a star can't support itself against gravity, it will collapse either to a white dwarf or a neutron star. Both objects are supported by quantum-mechanical forces.

1. The contraction heats the core enough to ignite hydrogen on its boundary, so that a hydrogen-burning shell develops. This usually is accompanied by an increase in luminosity due to the increased temperature.

2. The contracting core can reach a high enough temperature to make another reaction happen: the conversion of ^4He into ^{12}C . This reaction releases energy, and the star can settle for a while as it uses the helium as fuel. When this is exhausted, the core may contract, heat up further, and ignite carbon burning, converting ^{12}C into ^{16}O by the addition of ^4He . Reactions can go further, up to ^{56}Fe , but each step releases less energy and consequently lasts less time.

3. The release of an increasing amount of radiation from a contracting core actually makes the outer part of the star expand. The expansion can be very dramatic. Its radius can get so large that, despite its greatly increased luminosity, its surface temperature actually decreases, and it becomes distinctly red. Such a star is called a *red giant*. If the mass of the star is large (more than about eight times the mass of the Sun), we call the star a **supergiant**.

4. The star can get so large that gravity is very weak at its surface, and the pressure of the radiation leaving the star can blow off a steady strong "wind" of gas. Our Sun has such a wind, but giants have winds on much larger scales. They can lose large amounts of mass this way, up to 10^{-5} of a solar mass per year. On an astronomical time-scale, it does not take long for such winds to change the mass of the star significantly, or to transform the star into a shell of gas with a small hot star in the middle. Such shells are called **planetary nebulae**. Figure 12.5 on page 142 shows such a star. If a massive star loses enough mass, its hotter interior regions become exposed and it can become a blue giant.

The end of the main sequence lifetime of a star inevitably leads to great changes. These take place on relatively short time-scales, typically several million years. During this period, they are unusually bright, so that a large fraction of the stars visible to the naked eye are giants, even though they represent a small fraction of the total population of stars. The ultimate end of their post-main-sequence life is decided by how much mass the star has. That is the subject of the rest of this chapter.

Degenerate stars: what happens when the nuclear fire goes out

When a star's nuclear fuel runs out, which must eventually happen, then big change is unavoidable. Since gravity itself never "runs out", what happens next is dominated by gravity. In fact, it is easier to make calculations about the fate of a star than about its normal life. We shall see that relatively simple ideas lead us to white dwarfs, neutron stars and supernova explosions. We will continue the theme in Chapter 20 on neutron stars and Chapter 21 on black holes.

Investigation 12.2. Stars on the main sequence

This investigation is an exercise in extracting information from graphs, such as the curves in Figure 12.3 on page 139. This figure is plotted on what we call *logarithmic scales*. By that we mean that the main tick marks on, say, the vertical scale are not evenly spaced, but go up in powers of ten. The *logarithm* (the power of ten itself) of the vertical axis increases in uniform steps from one main tick mark to the next.

In this figure the horizontal axis is plotted logarithmically as well, although this is not so noticeable since the range is smaller. Whenever data span many powers of ten, and one is interested as much in the small values as in the big ones, it is useful to plot data this way.

What we see in Figure 12.3 is that all the curves are fairly straight. This implies something simple about the relations between quantities. A straight line has the general form of

$$y = mx + b, \quad (12.4)$$

where m is the slope of the line and b its y -intercept, the value of y where the line passes through the y -axis. If we consider the relation between luminosity L and stellar mass M in Figure 12.3, then the approximately straight line is a relation between their logarithms:

$$\log L = m \log M + b. \quad (12.5)$$

(Don't get confused between the slope m and the mass M : we use similar letters just in order to follow familiar notation for each.) We estimate m as follows. As M increases from 1 to 30, we see from the graph that L increases from 1 to about 1.5×10^5 . In terms of logarithms, $\log M$ increases from 0 to about 1.5 while $\log L$ increases from 0 to about 5.2. The slope is the change in $\log L$ divided by the change in $\log M$, or about 3.5.

Now, if we raise both sides of Equation 12.5 to the power of 10, we get a relation between L and M themselves. Recall the properties of logarithms:

$$10^{\log x} = x, \quad a \log x = \log(x^a), \quad \log x + \log y = \log(xy),$$

so that we get from Equation 12.5

$$L = \beta M^{3.5}, \quad (12.6)$$

where we define β by $\log \beta = b$. This is the way the y -intercept appears in the final result. We are more interested in the exponent

here (the slope of the curve in the figure) than in the intercept. We have learned that, for normal stars, luminosity is proportional to the 3.5-power of the star's mass.

In a similar way, we can deduce that the lifetime τ of a star is related to its mass by $\tau \propto M^{-2.5}$.

Since luminosity is energy released per unit time, the product of luminosity and lifetime is the total energy released over the star's life. If we then divide this by the mass of the star, we get the average energy released per kilogram. Given the proportionalities, we find

$$\frac{L\tau}{M} \propto \frac{M^{3.5} M^{-2.5}}{M} = \text{const.} \quad (12.7)$$

This is interesting because it is directly related to the nuclear reactions going on in the star. No matter what the total mass of the star is, each kilogram in it gives up on average about one million joules over the lifetime of the star.

This is what we called the "energy duty" in Investigation 11.1 on page 123 when we discussed where the Sun's energy came from. We concluded there that this energy was so large that it could only come from nuclear reactions. We now see that this is true for all stars, not just the Sun. Notice that this is the average energy given up by each kilogram of the star. Since most of the mass of any star is too far from the hot center for nuclear reactions to take place, the kilograms that actually do undergo nuclear reactions give up much more energy than this.

For the normal stars described in Figure 12.3 on page 139, the rapid increase in luminosity with mass has significant effects on the structure. Although there is more mass to generate gravity, there is much more pressure from the radiation flowing outwards through the star. The result is that more massive stars are larger. The radius of a normal star turns out to be roughly proportional to its mass to the 0.7 power: $R \propto M^{0.7}$.

Now, the surface temperature of a star is determined by its radius and luminosity: as we saw in Chapter 10, the luminosity is proportional to the fourth power of the temperature T and the square of the radius. Inverting this gives $T \propto L^{1/4} R^{-1/2}$. If we take the luminosity and radius to depend on the mass as above, we find two interesting relations: $T \propto M^{1/2}$ and $L \propto T^8$, where I have rounded off the exponents to simple integers and fractions.

Exercise 12.2.1: Inverting a logarithmic equation

Go through the steps leading from Equation 12.5 to Equation 12.6. First put the definition of β into Equation 12.5. Then write $m \log M$ as $\log(M^m)$. Finally combine this term with the β term on the right-hand side of Equation 12.5 to get the logarithm of Equation 12.6. Justify each step you make in terms of the rules given above for the use of logarithms.

Exercise 12.2.2: Dependence of stellar lifetime and luminosity on mass

In the same way as we estimated the exponent in the relationship between luminosity and mass, estimate the exponent in the relationship between lifetime and mass. Do you get -2.5, as given above? The curves for luminosity and lifetime are not perfect straight lines, so representing them by a single constant exponent is an approximation. Estimate the error in this approximation by giving a range of values for both exponents that would acceptably represent the graphs.

Exercise 12.2.3: Energy radiated per kilogram: is it a constant?

Test the assertion that the energy radiated per kilogram is constant, independent of the mass, by estimating the exponent from the graph in the same way as the lifetime and luminosity exponents were estimated. Is the exponent really zero? If not, can you explain this in terms of the uncertainty in the other two exponents that you arrived at in the previous exercise? In other words, is the exponent you get for the radiated energy within the range of exponents you would get if you selected various values for the other exponents and put them into the relation in Equation 12.7?

It would be natural to assume that, when the nuclear fuel runs out, the star will just begin to contract, and it will keep contracting until, perhaps, it shrinks to a point. After all, what can halt the contraction? What can provide as much resistance to gravity as had the now-exhausted nuclear reactions?

Remarkably, there *is* something, at least for small enough stars. This is the pressure of what we call a **degenerate gas**. To understand what a degenerate gas

is, we must go back to what we learned about the Heisenberg uncertainty principle in Chapter 7. If we try to pin down the location of, say, an electron to some great accuracy, then the *momentum* of the electron will be very uncertain. If the electron is confined in a certain region of space, say the volume of a star, then its position will be uncertain to no more than the size of the star. That creates a minimum uncertainty in its momentum, or equivalently in its velocity.

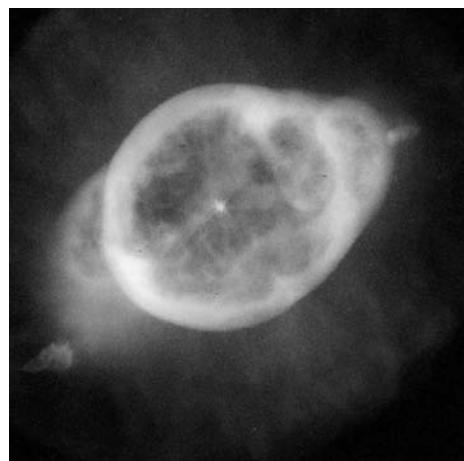


Figure 12.5. The planetary nebula NGC 3242: the glowing gas has been ejected by the star visible at the center of the nebula, and is now reflecting light from the star towards us. The complex and beautiful structure may result from the magnetic field of the star.
Image courtesy NASA/STSCI.

For ordinary stars, this is not important. But as a star contracts after the end of nuclear burning, the space in which its electrons are confined gets smaller, and so their minimum velocity goes up. This randomly oriented velocity provides an effective pressure that can act against gravity. In Investigation 12.3 we see that this can indeed counteract gravity, but on its own it would lead to a “star” of impossibly tiny dimensions, about 10^{-31} m in size! This is certainly not what actually happens, because there is another quantum principle at work: the *Pauli exclusion principle*, named for the Swiss nuclear physicist Wolfgang Pauli (1900–1958).

This says that electrons are individualistic: any electron will resist all attempts to force it to behave in an identical way to any other electron.

Now, electrons are intrinsically identical to one another. They all have exactly the same mass and electric charge. They also have another property, which comes purely from quantum theory: they have a small amount of angular momentum, called **spin**. You can think of each electron as spinning about some axis. The amount of spin is tiny, just $h/4\pi$. All particles in quantum theory have a spin (for some it is zero), and it always comes in multiples or half-multiples of the value $h/2\pi$: spin is quantized. The spin of an electron is half of this, and so electrons are said to be spin- $\frac{1}{2}$ particles.

Under the exclusion principle, no two electrons can be identical in all their properties. The only ways in which they can differ from one another is in their motion and spin. The remarkable aspect of spin- $\frac{1}{2}$ particles is that they can have only two *independent* orientations for their spin. Since no two electrons are allowed to have the same momentum and the spin, at most two electrons can have the same momentum.

The effect of the exclusion principle is that, as the star cools, only *two* electrons can have the minimum momentum allowed by the uncertainty principle. The rest of the 10^{57} electrons in the star have to have larger momentum, all of them different from each other by at least the minimum uncertainty in momentum. Such a gas is called a *degenerate electron gas*.

- ▷ Planck's constant h has dimensions of angular momentum.
- ▷ Spin is one of the deep mysteries of quantum physics. While it is possible to measure how an electron spins about any axis, there is a sense in which the axis of spin is not defined until it is measured. If two electrons have identical momentum and different spin, and if one of them is measured to be spinning about a certain axis in a clockwise sense, then the other one will always be measured to spin about the same direction in the counterclockwise sense. The second one acquires a direction of spin as soon as the first one is measured. We have no room in this book for a discussion of the fascinating subject of how measurements are made in quantum physics. That would require a book on its own!

The collapsing star will therefore have a population of electrons, some of which have quite a large momentum, hence quite a large energy. This population is called the *Fermi sea of electrons*, after the Italian nuclear physicist Enrico Fermi (1901–1954) who first described it. As we see in Investigation 12.4 on page 144, this Fermi sea will support the star against gravity in the manner described before, but because the electrons have much larger average momentum and energy, the balance between degeneracy pressure and gravity occurs when the star is much larger than if the ex-

Investigation 12.3. Degenerate matter, part 1: collapsing too far

Here we explore the physics of degenerate stars, where rather simple calculations lead us to striking conclusions. Our aim is to calculate how much pressure a degenerate star has. This pressure will be present even when a star is so cool that the ordinary thermal pressure does not hold it up. We will base our calculation on our study of the structure of the Sun and stars that we began in Chapter 8, and on the two fundamental principles of quantum theory that we have met so far: the Heisenberg uncertainty principle (Chapter 7) and the Pauli exclusion principle, discussed in the text of the present chapter.

In the uncertainty principle, the two related quantities we will use are the position of an electron on, say, the x -axis, and its x -momentum. Let an electron have speed v and momentum $m_e v$, where m_e is the mass of the electron. Suppose it is confined within a star of radius R . Since pressure comes from the momentum of particles, and since we want to find the pressure that is there even when thermal momentum has gone away, we want to find the *least* momentum allowed by the uncertainty principle. This will always be there.

The minimum momentum is associated with the maximum uncertainty in the position of an electron, which is the size of the star itself, $2R$. Assuming this, the x -momentum is at least

$$\Delta m_e v = \hbar / 2R, \quad (12.8)$$

where \hbar is Planck's constant. This contributes a kinetic energy of $(\Delta m_e v)^2 / 2m_e$, just from the x -motion. The three directions combine to give a *minimum kinetic energy* of

$$\langle KE \rangle_{\min} = 3\hbar^2 / 8m_e R^2. \quad (12.9)$$

This minimum random kinetic energy immediately leads us to the gas pressure. Recall our discussion of an ideal gas in Investigation 7.2 on page 78. The pressure of such a gas is given in terms of the random kinetic energy of the gas particles by Equation 7.7 on page 78. Substituting the above minimum kinetic energy into this gives a minimum pressure of

$$P_{\min} = \frac{\hbar^2 N_e}{4m_e V R^2}, \quad (12.10)$$

where V is the volume of the star and N_e is the number of electrons in the star.

Now, the structure of the star is a balance between pressure and gravity. We shall treat the structure equations approximately and see what we can learn from them about the degenerate star. We will then not expect the numerical values of quantities we deduce to be exact, but our results should represent at least roughly the relations between different physical quantities.

First we need the relation between the mass of the star and the number of electrons. The mass is determined by the number of protons and neutrons, since the electrons have negligible mass. If the star were composed of hydrogen, there would be only one proton per electron, and the mass of the star would be $m_p N_e$. More realistically, the star will be made of helium, carbon, and other elements up to iron. These typically have equal numbers of protons and neutrons

in their nuclei, so that the mass of the star is about $M = 2m_p N_e$. Although details of the composition of the star could change the factor of two by a small amount, our other approximations make more of a difference in the final answer than this.

Recall the equation giving the approximate way the pressure scales with the mass and size of a star, Equation 8.19 on page 95: $P = 3GM^2 / 4\pi R^4$. If we put the minimum pressure in here from Equation 12.10, use the mass we have just deduced, set the volume V to $4\pi R^3 / 3$, and solve for R , then we find the relatively simple relation

$$R = \frac{\hbar^2}{4G\mu m_e m_p M}. \quad (12.11)$$

If we take a solar mass for M (2×10^{30} kg) and use $\mu = 2$, then we have $R = 2.7 \times 10^{-31}$ m.

This is rather small! White dwarfs are indeed small stars: Sirius B is the size of the Earth. But here we have a radius much smaller than an atom! It is also much smaller than the radius we calculated for a black hole of a solar mass in Chapter 4, so we would not expect matter that behaves like this to form stars at all, but rather just to disappear into a black hole. So what has gone wrong with our calculation? The answer is that we have left something out of our considerations so far: the *Pauli exclusion principle*.

The Pauli exclusion principle applies to neutrons, protons, and electrons, which are the main constituents of matter. Such particles are called **fermions**. We will see in Investigation 12.4 on the following page that the exclusion principle requires fermions to form degenerate stars of the expected size.

We should note here, however, that the Pauli exclusion principle does not apply to all kinds of particles. Particles called **bosons** do not "exclude" one another. Photons are bosons, but they do not form stars by themselves because they do not have mass. Most other known bosons are unstable elementary particles, which would not last long enough to make a star. But there is speculation among particle physicists today that there may be a stable boson with a very small mass, and if it exists then it could in principle form **boson stars**, with radii smaller than white dwarfs. If bosons do not interact with one another, then the boson star would have a size given only by the uncertainty principle, as we have calculated above. The mass of the boson would then determine the maximum mass that a boson star could have without collapsing into a black hole. Conversely, in order to have a boson star of a certain mass, say a mass comparable to that of a white dwarf, then there is a maximum allowed mass for the bosons themselves. We examine this in the exercise below.

Larger boson stars are possible for a given boson mass if bosons repel one another, so that the uncertainty principle does not determine the size of the star. At the present time (2002) there is no compelling experimental evidence that such particles or stars might exist, but the Universe has proved itself to be full of surprises. Such stars could be detected by observing the gravitational radiation emitted by binary pairs of boson stars. As we shall see in Chapter 22, gravitational wave detectors will routinely conduct searches that could reveal such systems.

Exercise 12.3.1: Boson stars

- (a) In Equation 12.11, replace the proton and electron masses by a single boson mass m_b and assume that $\mu = 1$. This gives the formula for a star composed of just one type a particle, the boson of mass m_b . For such a star of total mass M , calculate the following ratio

$$2GM/Rc^2 = 8M^2 m_b^2 / m_{\text{Pl}}^4,$$

where m_{Pl} is the Planck mass defined in Equation 12.20 on page 146.

(b) The ratio above is the ratio of the size of a black hole, $2GM/c^2$, as given in Equation 4.12 on page 36, to the size of the star. We will see in Chapter 21 that the star cannot be smaller than a black hole, so this ratio must be less than one. Show that this sets a maximum mass on a boson star made from bosons of mass m_b :

$$M_{\max} = 8^{-1/2} m_{\text{Pl}}^2 / m_b.$$

- (c) Find the largest mass m_b that the boson could have in order to allow boson stars of a solar mass to exist. Find the ratio of this mass to the mass of a proton. You should find that the boson needs to have very much less mass than a proton.

Investigation 12.4. Degenerate matter, part 2: white dwarfs

According to the Pauli exclusion principle, if there are N_e electrons, they all must form pairs that have different values of the momentum separated by at least $\Delta m_e v$. It follows that, in three dimensions, the largest momentum in any direction must be at least $N_e^{1/3} \Delta m_e v$. (We omit factors of 2 and π and so on here, since our treatment of the structure of the star is approximate anyway. If you are unsure of where the factor of $N_e^{1/3}$ comes from, see Exercise 12.4.1 below for a derivation.) The average momentum is within a factor of two of this, so we will take this to be the typical momentum of an electron in a degenerate Fermi gas.

Going through the steps leading to Equation 12.11 on the previous page again gives this time a radius larger by a factor of $N_e^{2/3}$:

$$R = N_e^{2/3} \frac{h^2}{4G\mu m_e m_p M}.$$

Using the fact that $N_e = M/\mu m_p$, we have

$$R = \frac{h^2}{4Gm_e(\mu m_p)^{5/3} M^{1/3}} = 10^7 \text{ m, for } M = 1 M_\odot. \quad (12.12)$$

This radius is slightly greater than the radius of the Earth. As we noted above, we can trust it to be correct to perhaps a factor of two or so, but not better than that. In fact, more detailed calculations show that a white dwarf of one solar mass has a radius about 90% of that of the Earth. Our answer is quite close to the right one, considering the simplicity of the mathematical approximations we have made. This tells us that we have not left out important physics.

Exercise 12.4.1: Momentum in the Fermi sea

Here is where the factor of $N_e^{1/3}$ comes from. First we consider the easier case of electrons confined in a one-dimensional “box”, say along a string of finite length. We return to the three-dimensional star later. If each pair of electrons has a distinct momentum, separated by $\Delta m_e v$ from its neighbors, then we could mark out a line on a piece of paper, start with the smallest momentum allowed ($\Delta m_e v$), and make a mark each step of $\Delta m_e v$. Each mark represents the momentum of one pair of electrons. If we have N_e electrons, then there will be a total of $N_e/2$ marks. We would have to make marks in the negative direction too (electrons moving to the left), so the largest momentum will be $(N_e/4)\Delta m_e v$. Now suppose the electrons are confined to a two-dimensional square sheet of paper. Show that, leaving out factors of order unity, their maximum momentum is $N_e^{1/2} \Delta m_e v$. (Hint: each pair of electrons occupies a square of momentum uncertainty.) Similarly, show for three dimensions that the result is $N_e^{1/3} \Delta m_e v$.

What is the *equation of state* of a degenerate Fermi gas? Recall our use in Investigation 8.4 on page 92 of a power-law (polytropic) relation between pressure and density to make a simple model of the Sun. Here we will find that the degenerate Fermi gas obeys such a law.

The pressure given in Equation 12.10 on the previous page for a gas without the exclusion principle needs to be multiplied by $N_e^{2/3}$, the square of the factor by which the average momentum goes up when we take account of the exclusion principle. That gives

$$p_{\text{Fermi}} = \frac{h^2 N_e^{5/3}}{4m_e V R^2}. \quad (12.13)$$

If we note that $R \propto V^{1/3}$, we see that the pressure depends only on the ratio N_e/V , the number of electrons per unit volume. Since this is itself proportional to the mass density $\rho = \mu m_p N_e/V$, we arrive at the *Fermi equation of state*

$$p_{\text{Fermi}} = \beta \rho^{5/3}, \quad (12.14)$$

where β is a constant that depends on h , m_e , μ and m_p . Our value for β is not exact because our calculations are approximate in places, but what is exact is the exponent: the degenerate Fermi gas is a polytrope whose polytropic index, as defined by Equation 8.12 on page 92, is $n = 3/2$.

clusion principle were not operating. A star supported by electron degeneracy is called a *white dwarf*. We show in Investigation 12.4 that a white dwarf’s radius depends on its mass and on a simple combination of fundamental constants of physics (specifically, on G , m_e , m_p , and h). A star with the mass of the Sun should be about the same size as the Earth. Its density is, therefore, huge: about one million times the density of water!

The protons in the star should also be subject to the same uncertainty and exclusion principles, and therefore to the same minimum momentum. This will create a degenerate proton gas and an associated degeneracy pressure. But the kinetic energy of a proton that has the same momentum as an electron is much less than that of the electron: kinetic energy is $\frac{1}{2}mv^2 = (mv)^2/2m$, and since momentum is just mv , it follows that the kinetic energy of two particles that have the same momentum is inversely proportional to their masses. The proton is nearly 2000 times more massive than the electron, so the degenerate proton gas would have only $1/2000^{\text{th}}$ of the energy of the degenerate electron gas. For this reason, the electrons provide essentially all the pressure in such a situation.

Not all particles are subject to the exclusion principle. In fact, the exclusion principle only applies to particles with half-integer spin, such as spin- $\frac{1}{2}$. This includes all the ordinary particles of matter: electrons, protons, and neutrons. But some particles have whole-integer spin, either spin-0 (pions), spin-1 (photons), or – as we will see in Chapter 27 – spin-2 (gravitons). These do not exclude one another; in fact they have some preference for ganging up together in the same state! Photons,

►Pions are elementary particles that are made in particle accelerator experiments. Gravitons are the quantized form of gravitational waves.

for example, would never form the kind of electromagnetic waves that radios and mobile phones use if no two of them could be the same. Lasers, which essentially emit strong beams of light with all the photons in exactly the same wavelength and state of oscillation, exist only because photons like being the same as one another. Particles that obey the exclusion principle are called fermions, after Fermi. Particles that like being together are called bosons, after the Indian physicist Satyendra Nath Bose (1854–1948).

Although degenerate matter has strange properties, the natural evolution of a star brings it to the point where degeneracy becomes important. Astronomers see white dwarfs in their telescopes, and they have just the size we have calculated. These huge objects depend for their very existence on the strange physics of quantum theory. Because they are composed of matter whose structure cannot be described in the conventional language of forces, they would have been incomprehensible to Newton. Yet they are abundant: one in ten stars is a white dwarf; and, as we noted in Chapter 10, one of the brightest stars in the night sky, Sirius, has a white dwarf in orbit about it.

The Chandrasekhar mass: white dwarfs can't get too heavy

Unfortunately for some stars, but fortunately for the evolution of life on Earth, the story of degeneracy does not stop here. The problem is that, if the star is very massive, then gravity will force the degenerate electrons into such a small volume that their typical speed becomes close to the speed of light. In this case, we have to treat the electrons by the rules of special relativity.

We will study special relativity beginning in Chapter 15, but here we need only one new fact that is explained there: the momentum carried by a photon is just proportional to its energy, in fact is E/c . This is very different from the situation for low-speed electrons, where the momentum is twice the kinetic energy divided by the speed of the particle. The difference is in part due to the fact that in special relativity, mass has energy, so the relationship between energy and momentum must include the total mass-energy of a particle. Now, since at a speed close enough to the speed of light, any particle behaves more and more like a photon, the momentum carried by a fast electron is also just its energy divided by c .

Now, the energy of the gas particles is directly responsible for the pressure of the gas, so when the electrons become relativistic, the pressure starts to increase only in proportion to the uncertainty momentum, rather than to its square. This significantly weakens the degeneracy pressure that electrons can exert, and in fact we show in Investigation 12.5 on the next page that it leads to a universal *maximum mass of a white dwarf*, the maximum mass that can be supported by degenerate electrons. This mass is called the *Chandrasekhar mass* M_{Ch} , after the Indian astrophysicist Subrahmanyan Chandrasekhar (1910–1995) who discovered it. Its value is in the range 1.2 to 1.4 times the mass of the Sun, depending on the exact composition of the star when it reaches the density of the white dwarf.

The Chandrasekhar mass is one of the most remarkable numbers in all of physics. As we show in Investigation 12.5 on the following page, it depends mainly on some fundamental constants of nature, not on fine details of atomic or nuclear physics.

It seems to be an accident of our Universe that this particular combination of the constants of nature gives a mass for a relativistic degenerate white dwarf that is similar to the masses of ordinary stars. If this mass had come out to be much larger or much smaller than a solar mass, the death of most stars would be radically

In this section: the kind of quantum-mechanical support that white dwarfs use can only support a little more than the mass of the Sun. With more mass, the star will collapse.



Figure 12.6. Subrahmanyan Chandrasekhar made many important contributions to astrophysics, sharing the 1983 Nobel Prize for Physics. His discovery as a very young man of the limiting mass for white dwarfs led to a bitter conflict with Eddington (see Chapter 4), who felt that Nature simply could not behave in such a way! Chandrasekhar lost this battle, escaping from Cambridge to Chicago. Subsequent research, of course, vindicated his work completely. Chandrasekhar's modest manner, his devotion to science, and his erudition won him the respect and affection of generations of scientists. Image courtesy University of Chicago.

Investigation 12.5. Deriving the Chandrasekhar Mass

For stars with relativistic electrons, we need to re-calculate the structure and equation of state from the beginning, since Equation 12.9 on page 143 is wrong for this case. As we will see in Chapter 15, the energy of a photon is just the speed of light times its momentum. Therefore, this must be almost true even for ordinary particles moving at close to the speed of light:

$$E = pc. \quad (12.15)$$

Given the same uncertainty in momentum, $\Delta m_e v = h/2R$, using the exclusion principle to give a typical momentum that is a factor of $N_e^{1/3}$ larger than this, and then following the same steps as before for the ideal gas, we arrive at a pressure

$$p = \frac{hcN_e^{4/3}}{3RV}. \quad (12.16)$$

Setting this equal to $3GM^2/4\pi R^4$ as in Investigation 12.3 on page 143, and replacing the volume by $4\pi R^3/3$, we get

$$\frac{hcN_e^{4/3}}{4\pi R^4} = \frac{3GM^2}{4\pi R^4}. \quad (12.17)$$

Here we notice a remarkable and unexpected thing: the radius of the star drops out, and we are left with an equation that determines a single mass! This mass is the *unique* mass of a fully relativistic white dwarf. For non-relativistic white dwarfs we could choose a mass and find a radius, or vice versa. Here, we have no choice about the mass, and presumably the radius can be anything at all! This rather remarkable discovery was made by Chandrasekhar, and so we name the unique mass after him. Our expression for it is, from the previous equation,

$$M_{Ch} = \left(\frac{hc}{3G} \right)^{3/2} \left(\frac{1}{\mu m_p} \right)^2. \quad (12.18)$$

This evaluates to about 1.4 solar masses. Of course, our calculation is only approximate, but it turns out that we have got the right value almost exactly.

Since real electrons don't exactly obey Equation 12.15, but come closer and closer to it the more relativistic they get, we should regard the Chandrasekhar mass also not as the exact mass of any particular white dwarf but rather as an upper bound on the mass of all white dwarfs. Less massive stars have fewer relativistic electrons. More massive stars simply cannot be supported by electron degeneracy pressure at all.

Is the fact that we have not determined the radius of this star a worry? Not really: again, in a real star, the electrons are not fully

relativistic, the electron gas is not perfectly ideal, and there is some pressure support from the protons or other nuclei. All these make small corrections, but they are enough to guarantee that any real star's radius will be determined by its mass. The radius will be about the radius of the Earth, as before.

Importantly, the equation of state of the relativistic white dwarf is also a polytrope, but this time with a different power. Steps similar to those used in Investigation 12.3 on page 143 give the relation

$$p = \beta \rho^{4/3}, \quad (12.19)$$

so that the polytropic index is 3.

Now we remind ourselves of the calculation we did for the stability of the Sun, in Investigation 8.8 on page 101. A polytrope of index 3 is only marginally stable against collapse. Any small correction to the properties of white dwarfs could cause them to be unstable. One correction is that, on compression, some electrons and protons tend to combine into neutrons, removing electrons from the degenerate sea and reducing its pressure. This makes collapse more likely. A second correction is general relativity: in Einstein's theory of gravity, the critical polytropic index actually needs to be somewhat smaller than 3, so that an $n = 3$ gas is actually unstable. Both of these effects become important for highly relativistic white dwarfs, and lead them to be unstable to gravitational collapse a bit before they reach the Chandrasekhar mass.

It is interesting to note that the Chandrasekhar mass can be expressed in terms of two simpler masses: the proton mass m_p and a number with the dimensions of mass that is built only out of the fundamental constants of physics, h , c , and G :

$$m_{Pl} = \left(\frac{hc}{G} \right)^{1/2} = 5.5 \times 10^{-8} \text{ kg}. \quad (12.20)$$

This mass is called the **Planck mass**, hence the symbol m_{Pl} . In terms of these simple masses we have

$$M_{Ch} = \left(\frac{1}{3^{3/2} \mu^2} \right) \frac{m_{Pl}^3}{m_p^2}. \quad (12.21)$$

The Planck mass was first discussed by Planck himself. He noticed, soon after introducing his constant h , that from the fundamental constants h , c , and G one could build numbers with any dimensions one wanted: a mass, a length, a time, and so on. These are now called the Planck mass, the Planck length, etc. We do not yet know exactly what role these quantities play in physics, but we expect it to be fundamentally associated with the quantization of gravity, since they involve both h and G . We will return to this in Chapter 21.

Exercise 12.5.1: Deriving the Chandrasekhar mass

Derive the expression in Equation 12.18 by the indicated method.

Exercise 12.5.2: Relativistic degenerate gas equation of state

Find the constant β in Equation 12.19.

different. Since stellar death provided our Solar System with the raw ingredients of life, we owe much to the Chandrasekhar mass!

Neutron stars

What, then, happens to a contracting star if its mass exceeds the Chandrasekhar mass? Electron degeneracy fails because the electrons have become relativistic, but the protons are still available. Because the proton mass is much larger than that of the electron, protons do not become relativistic until the star is much smaller. When the star is the size of a white dwarf the proton degeneracy pressure is negligible, but as the star contracts further this pressure grows until it can support the star.

The calculations in Investigation 12.4 on page 144 show that a degenerate star's

In this section: neutron stars are also supported by degeneracy pressure, but here it is the neutrons which form the supporting distribution. Their maximum mass exceeds $2M_\odot$.

radius is inversely proportional to the mass of the particle that is responsible for the degeneracy pressure, and it also depends somewhat on the composition. As we show in Investigation 12.6 on the next page, the result is that the star will continue to contract until it reaches a radius about 1/600th of the radius of a white dwarf. This size is about 10–20 km.

Can there really be stars with the mass of the Sun that are only 10 km in size?? Remarkably, the predictions of our simple calculations are borne out by observations: more than 1000 such *neutron stars* have now been observed, and possibly about 0.1% of all stars are this size!

In order to understand how we can identify such neutron stars in astronomical observations, we need to study them in more detail.

While the contraction from the white dwarf stage is occurring, a crucial change takes place within the material of the star. As the density increases, the energy of a typical degenerate electron gets to be so large that it exceeds the energy equivalent of the difference between the mass of the proton and the neutron. (The neutron is slightly more massive than the proton.) The result is that it costs less energy to combine a proton and an electron into a neutron than it does to keep the electron in the Fermi sea. The electrons then almost all combine with protons to form neutrons. The contracting material is then a *degenerate neutron gas*. The energy that is released by this nuclear transformation is carried away by the neutrinos that are given off when this happens.

This transformation makes no difference to the degeneracy pressure, since neutrons have essentially the same mass as protons, and they both, like electrons, obey the Pauli exclusion principle. Moreover, as we see in Investigation 12.6 on the following page, the maximum mass that can be supported by neutron degeneracy pressure is much larger than the Chandrasekhar mass for white dwarfs, so, unless the contracting star is very massive, the contraction of the star can be halted when neutrons become degenerate.

This is why we call this a *neutron star*. As we noted, its radius is about 10–20 km. We shall study the properties of these ultradense objects in some detail in Chapter 20, and learn there that they are associated with *pulsars*, of which some 1000 have now been identified.

It is worth noting here, however, that neutron stars take us into the province of relativity, where we should not trust Newtonian gravity too much. We can see this by calculating the escape velocity from a neutron star whose mass is $M = 1M_{\odot}$ and whose radius is $R = 10$ km. The Newtonian formula for the escape velocity is $v^2 = 2GM/R$. A little arithmetic gives $v = 1.6 \times 10^8$ m s⁻¹, or about half the speed of light! Newtonian gravity can only indicate the general features of neutron stars, but it cannot give a good quantitative description of them. That will have to wait until Chapter 20.

Fire or ice: supernova or white dwarf

We now have enough understanding of the possible forms of equilibrium stars to look at what happens to giant stars at the end of their nuclear lives. Some die away quietly as white dwarfs, and others explode spectacularly as supernovae.

We saw above that giants have cores made of the waste products of nuclear reactions. These waste products are nuclei of moderate mass, like carbon and oxygen, mixed with helium. There is no hydrogen in the core, of course. The core is surrounded by a shell in which nuclear reactions still take place, but as the material in the shell is exhausted, the shell moves outwards and the core increases in mass. This burned-out core is supported by electron degeneracy pressure – there is no

In this section: when a star runs out of nuclear fuel, it can end in a supernova explosion or in a quieter contraction to a white dwarf.

*Some say the world will end in fire,
some say in ice, ...*

(Robert Frost)

Investigation 12.6. Neutron stars as degenerate stars

Now, the degeneracy calculations we have performed in previous investigations have not used any special properties of the electrons: they are just the particles that supply the degeneracy pressure. As we note in the text, a *neutron star* is a star where neutrons supply the degeneracy pressure. All the formulas before apply then to neutron stars, if we replace m_e by m_n , which to our accuracy is the same as m_p . We can also set $\mu = 1$, since practically all the particles in a neutron star are neutrons. This means that the radius of the star from Equation 12.12 on page 144 will be smaller by the ratio m_e/m_p but larger by the fact that the white dwarf factor of $\mu^{5/3} = 3.2$ is set to one for neutron stars. This means the radius of a neutron star

should be about 1/600th of that of a white dwarf, or about 17 km. This is within the range of what the more detailed calculations give, even in general relativity (see Chapter 20).

Notice also that the Chandrasekhar mass in Investigation 12.5 on page 146 for neutron stars should be larger than that for white dwarfs only because of the factor of μ^2 . This raises the maximum mass of a neutron star to about five or six solar masses. The effects of relativity, however, drastically reduce this number. We shall see in Chapter 20 that the extra strength of relativistic gravity reduces the maximum mass of a neutron star to about $2M_\odot$.

other available form of support – and so it is just a small white dwarf inside the giant. The composition of its nuclei has little effect on the electron pressure, but it has a small effect on the size of the Chandrasekhar mass: stars with heavier nuclei have a larger critical mass. What happens as the core grows depends on the initial mass of the star. It is a complex process that astrophysicists must simulate on computers in order to understand. The outline of what happens is clear now, but many details are still poorly understood.

For stars of moderate mass, say less than about 8–10 solar masses, the growth of the core is slow enough, and the stellar wind at the surface of the giant is large enough, that before the mass of the core reaches the Chandrasekhar mass the rest of the star has been blown away by the steady wind. The core never exceeds the Chandrasekhar mass, and the endpoint is a white dwarf star that gradually cools off, a dead degenerate cinder. Our own Sun, which starts out with less than a Chandrasekhar mass, will end like this, as a white dwarf of perhaps 0.6 solar masses. For a time the expanding shell of expelled gas is be illuminated by the part of the star that still remains, forming a beautiful planetary nebula, as in Figure 12.5 on page 142.

This peaceful and silent end is not available to more massive stars. Above about 8–10 solar masses, the loss of mass during the giant phase is not fast enough to prevent the core reaching the Chandrasekhar mass. When this happens, it collapses.

The core that collapses does not have the original composition of the waste products. Typically by this time it has changed and become dominated by iron. We saw in the previous chapter that iron is the natural end-point of nuclear reactions, that reactions among carbon and oxygen and other nuclei release energy when they form iron. As the core of the giant accumulates mass, these reactions convert most of the nuclei to iron. So by the time the core reaches the Chandrasekhar mass and collapses, its composition is inert. It has no chance to release energy from further reactions during its collapse.

There is therefore nothing now to halt the collapse until the former white dwarf reaches neutron star density, where neutron degeneracy pressure can build up. During the collapse, the increasing density makes the protons in the iron nuclei combine with the free electrons to form neutrons. When the collapse reaches the density of a neutron star, the speed of infall will be nearly the escape velocity from a neutron star, since the starting point of the collapse is very large compared to the size of a neutron star. We estimated above that this is about half the speed of light! At this speed, the collapsing star shrinks from white dwarf size (6×10^6 m) to neutron star size in less than a tenth of a second. We call this free fall **gravitational collapse**.

What happens when neutron degeneracy pressure builds up to halt this incredible speed? If there were no friction, no dissipation of the energy of collapse, then the star would have to “bounce” and re-expand, just like a rubber ball hitting a brick wall. However, the conversion of electrons and protons into neutrons dur-

▷ The wind from a giant is much stronger than the wind that comes from our Sun at present. Astrophysicists do not completely understand the complex processes that lead to the expulsion of such a large amount of material.

▷ The situation is like that of Investigation 11.1 on page 123, where a star radiates its thermal energy away. White dwarfs have only 10^{-4} of the surface area of the Sun, making their luminosity is smaller by the same factor. It can therefore take tens of billions of years for the dwarf to cool off.

▷ Why don't these reactions go rapidly? After all, they release energy, so they need no energy input to drive them. The barrier is the electric repulsion of nuclei for each other. At low temperatures the nuclei have low speeds and don't come near enough to one another to trigger nuclear reactions. Only when the temperature reaches a certain value do nuclei have enough speed to overcome their electric repulsion and begin to react. This temperature is reached as the core grows towards the Chandrasekhar mass inside a massive giant star.

ing the collapse has produced neutrinos that have already removed some energy, so the bounce will be a little weaker. What is more, at the point of maximum density something new happens. Neutrinos, which can pass through normal matter virtually without scattering, become trapped: the density is high enough to scatter them many times as they move through the star.

The effect of this is that neutrinos quickly come into thermal equilibrium with the hot neutron matter, and a neutrino gas builds up. Much of the kinetic energy of infall is converted into neutrino energy. When the bounce starts and the density goes down a little, these neutrinos can suddenly escape, carrying away a great deal of energy. This is a sort of shock absorber, which prevents the star from rebounding back to its original white dwarf size again. Most of the star is now trapped at the enormous density of the neutron star.

But the rest of the material of the star has also been falling in, and begins to hit the outer layers of the core, just as the neutrinos are beginning to expand away. The neutrino gas runs into the infalling envelope of the star, and what physicists call a "shock wave" develops. Familiar examples of shock waves are the sonic boom, the bow wave in the water in front of a fast-moving ship, and the tidal bore found on some rivers, as in Figure 5.4 on page 44.

What happens after the shock forms seems to depend sensitively on details of the nuclear physics, much of which is not yet fully understood. But computer simulations suggest that, at least in many cases, the neutrinos remain trapped long enough to help push the shock outwards into the infalling envelope, with enough energy to blow the envelope away. The expanding envelope is heated by the shock, so that nuclear reactions take place in it at a very rapid rate. When the shock reaches the outer boundary of the star, the star suddenly brightens up, and we see a *supernova*. Meanwhile, at the center, the collapsed core either settles down into a neutron star, or – if much further material from the envelope falls down onto it – collapses again to a black hole. We will discuss both possible outcomes in later chapters.

It should not be a surprise that the envelope can be blown away by the neutrinos. The energy released by the collapse of the core is enormous. When we study general relativity later in this book we will learn how gravity can convert mass into energy. Gravitational collapse to a neutron star converts a larger fraction of the mass of the core into energy than happens in a nuclear reactor or nuclear bomb, and much of this energy is carried away by the shock. The envelope has been sitting in a relatively weak Newtonian gravitational field, and it is no match for the thundering impact of the shock. Despite the fact that the envelope may contain ten or twenty times the mass of the core, it blows away at a high speed.

What we have described is called by astronomers a **supernova of Type II**. Supernovae of Type II are among the most spectacular events visible in optical telescopes. The most recent one visible to the naked eye from the Earth was the supernova of 1987, called SN1987A (see Figure 12.7). Located in the Large Magellanic Cloud, which is a small galaxy in orbit about our own Milky Way galaxy (see Chapter 14), it seems to have occurred in a blue giant star of about 20 solar masses. It

►A shock develops when an object (here the expanding core) moves into a fluid (the envelope) with a speed faster than the local sound speed. The fluid cannot move out of the way fast enough and a large density difference develops just in front of the object.

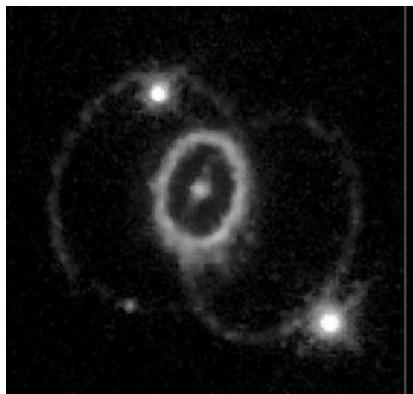


Figure 12.7. The supernova of 1987 in the Large Magellanic Cloud was accompanied by the formation of these extraordinary rings when light from the supernova hit shells of gas that the original giant star had expelled during a phase of mass-loss. The supernova light caused these shells to glow in fluorescence. Image courtesy of NASA/STSCI.

►We will discuss supernovae of Type I below.

Figure 12.8. The Crab Nebula is the result of a supernova explosion recorded by Chinese astronomers in 1054. The explosion left behind a neutron star, which today is seen as a pulsar (see Chapter 20 and especially Figure 20.4 on page 270).

The pulsar is the lower of the two bright stars near the center of the nebula, oriented along a diagonal line from lower right to upper left.

This nebula is also known as M1.

Photo by Jay Gallagher (U. Wisconsin)/WIYN/NOAO/NSF.



was the brightest star in the southern hemisphere sky for a time.

Physicists were fortunate enough to have detected not only light but also the neutrinos from this supernova: when the trapped neutrinos escaped, about 11 of them induced nuclear reactions in the Kamiokande proton-decay experiment in Japan, and a few others registered in similar experiments in the USA and Russia. This number is about what one would have expected, and their detection provided a clear verification that the picture of the supernova mechanism described here is fundamentally right.

Supernova explosions are not commonplace events. The last one visible to the naked eye before SN1987A was recorded by Kepler in 1604. Once the supernova is triggered, the cloud of gas continues to expand for thousands of years. Astronomers see many such supernova remnants relatively near the Sun. The most spectacular is the Crab Nebula, shown in Figure 12.8.

We are particularly fortunate to be able to see in this remnant the neutron star that was formed by the explosion. This neutron star is a *pulsar*, and we shall discuss it in Chapter 20.

Death by disintegration

In this section: some stars end in a giant nuclear explosion. These are called Type I supernovae.

A supernova of Type II leaves a neutron star or black hole behind. Other explosions can be so violent that they leave nothing behind at all. These occur when a white dwarf star formed long before undergoes a long-delayed gravitational collapse. This

can happen when, for example, the star is in a binary system and gas from the companion star falls onto the white dwarf. We will discuss such phenomena in the next chapter.

After a long period of accumulating mass, the old white dwarf could reach the Chandrasekhar mass. What happens next is very different from what happens inside a giant. The old white dwarf is still composed of carbon, oxygen, and other nuclei lighter than iron. It is hot, but not hot enough to allow the nuclear reactions that form iron to take place. The material falling on it is mostly hydrogen, which converts quickly to helium at the temperature of a white dwarf, but not to iron. So there is still plenty of nuclear energy available in the material of an old white dwarf.

When its long-postponed collapse finally begins, the old white dwarf is an enormous nuclear bomb waiting to happen.

During the collapse the increasing density and pressure leads to a rapid increase in nuclear reactions, as the nuclei combine to form heavier nuclei. Since most of the original nuclei are lighter than iron, these reactions release energy. This energy is enough to stop the collapse well before it reaches the density of a neutron star and blow the white dwarf completely apart. This is what most astronomers believe is the mechanism of the explosion that they call a **supernova of Type Ia**. These explosions are the brightest of all supernovae, and astronomers have been able to detect them at huge distances. In fact, although Type Ia supernovae are even rarer than Type II, astronomers have been able to find enough of them at very great distances to demonstrate that the expansion of the Universe is apparently accelerating rather than slowing down. We will come back to this extraordinary and unexpected observation in Chapter 14 and in the final chapters on cosmology.

What is left behind: cinders and seeds

When almost all stars die, much of their material gets locked up forever in a stellar cinder: a white dwarf, a neutron star, or even a black hole. But at the same time, much of their material is returned to interstellar space to be recycled into further generations of stars. Winds that blow away the outer envelopes of massive stars return gas that is enriched in carbon, oxygen, and other elements vital to life. The gas ejected in a supernova explosion is different. It has been thoroughly processed by the shock wave that ejected it, and the material ejected is rich in very heavy elements, from iron to uranium. Not only are some of these elements vital for life, but as we have seen the uranium powers geological activity on the Earth, without which life could not have evolved. Without white dwarf cores to trigger supernovae and neutron stars to stop the collapse and generate the bounce, we would not be here!

►Nuclear reactions do not release as much energy as gravitational collapse, so a Type II supernova is more energetic than a Type Ia. But most of this energy comes out as kinetic energy of the envelope, so the photon luminosity of a Type Ia is in fact larger than that of a Type II.

In this section: we remind ourselves that the existence of compact star remnants is essential for the formation of life.

Binary stars: tidal forces on a huge scale

Binary stars are stars bound in orbit about one another by their gravitational attraction. Most stars seem to form in binary systems or in systems containing more than two stars. This is not really surprising: stars form from condensations in giant clouds of gas, so where one star forms, others are likely, and they may form close enough to each other to be bound together forever. We saw this in the numerical simulation reproduced in Figure 12.2 on page 138.

We have already studied special cases of binaries: planets in motion around the Sun, and the Moon around the Earth. These orbits allow us to measure masses in the Solar System. We learn the Sun's mass once we know the radius and period of the Earth's orbit. Similarly, we measure the mass of the Earth by studying the motion of the Moon (and of artificial Earth satellites). In the same way, binary star orbits are used to measure the stars' masses. Binaries are often our best, indeed our only, way of measuring the masses of stars. When two compact stars (such as neutron stars or black holes) form a binary system, they can be used to test our ideas about gravity: one of the most stringent tests of general relativity is that it predicts perfectly the observed orbits in a certain binary neutron star system.

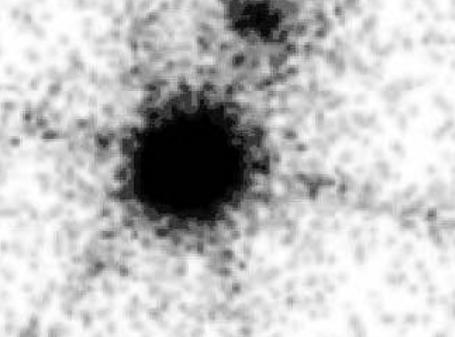
The nuclear physics in the core of a star in a binary system is normally not affected much by the companion. The core is fairly dense and small, so the tidal forces on it from the other star do not change the conditions there very much. But the gravitational field of the companion can have dramatic effects on the outer parts of the star, and these can eventually change completely the course of evolution of the star. When binaries consist of a main sequence star and a cinder (a neutron star, for example), the cinder can strip the ordinary star of gas, with spectacular results. Such binaries are sites where nova explosions occur; they can produce X-rays in abundance; and they sometimes shoot out extraordinarily narrow beams of particles traveling at nearly the speed of light. Binary evolution is one of the most intensively studied subjects in astronomy today.

Looking at binaries

Binary systems turn up in astronomical observations in many different ways. Some binaries are easy to observe directly: with a telescope one can watch, over a period of months or years, the positions of two stars change as they orbit one another. Such **visual binaries** are relatively rare, since they have to be close enough to us for our telescopes to be able to separate the two images despite the blurring caused by the Earth's atmosphere. Sirius A and B, described in Chapter 10, are a good example of a visual binary. The Hipparcos satellite, mentioned in Chapter 9, has greatly increased the number of such binaries known.

It is much more common to learn that a star is in a binary system by recording its *spectrum*. We saw in Figure 10.3 on page 114 that the spectrum of a star is peppered with sharp features called spectral lines. The wavelengths of these lines are characteristic of the atoms that emit or absorb the light. These wavelengths

In this chapter: we look at a number of astronomical systems that are affected by tidal forces, inhomogeneity of the gravitational field. These systems include binary stars, interactions between planets, mass flows between stars, X-ray binaries, and the three-body problem. We use computer simulations to explore realistic examples of many of these systems.



►The background picture on this page is an X-ray image of the binary stars Sirius A and B, taken by the Chandra satellite. The rays are optical effects produced by the telescope. The Sirius system was discussed in Chapter 10. Image courtesy NASA/SAO/cxc.

In this section: we learn how astronomers know which stars form binary systems, and what can be learned from observations.

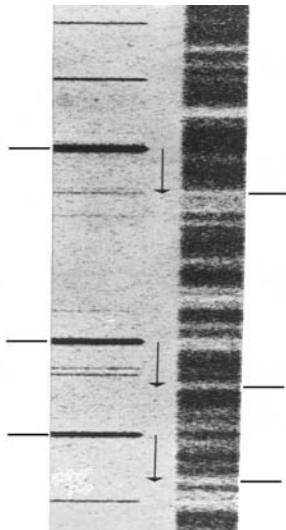


Figure 13.1. The spectral lines of the observed spectrum (left, in negative) are shifted from those of the reference spectrum (right). The amount of the shift indicates the velocity of the star along the line-of-sight. The marked lines are all lines of neutral iron.

The amount of the shift indicates the velocity of the star along the line-of-sight. The marked lines are all lines of neutral iron.

can be measured in the laboratory and compared with the observed spectrum. If the star is moving away from us, then the Doppler shift, explained in Figure 2.3 on page 15, will shift the wavelengths of all the lines to the red (longer wavelengths). Similarly, if the star is moving towards us, the lines shift to the blue. So by comparing the positions of spectral lines in an observed spectrum with those in the lab, we can determine the speed of the star along the line-of-sight. This is illustrated in Figure 13.1. Importantly, motion of the star in directions perpendicular to the line-of-sight produces no Doppler effects and is not measurable this way.

If we look at the spectrum of a star in a binary system, then the Doppler shift of the lines should change with time, as the star's orbital velocity changes. Observed for long enough, the changes should be periodic, that is they should repeat after one orbital period of the binary. Binaries that are discovered this way are called **spectroscopic binaries**, and they constitute the overwhelming majority of known binaries.

If both stars are of comparable brightness, then it may be possible to see two sets of spectral lines, shifting in different ways but with the same period. But it usually happens that only one set of lines is visible, either because the second star is too dim for its lines to be seen in the light from the first, or because the second star is a "cinder" (a white dwarf or a neutron star) that does not have prominent spectral lines.

In a spectroscopic binary, we only learn about velocities along the line-of-sight. The orbital plane, on the other hand, will be oriented at random to the line-of-sight. So we only get partial information about the orbit. In an extreme case, if we happen to be looking at the orbit "face-on", directly down onto the orbital plane, there will be no motions along the line-of-sight, no Doppler shifts, and we might not recognize the system as a binary at all. At the other extreme, if our line-of-sight is in the plane of the orbit, we see the whole motion. Astronomers who wish to use binaries to measure the masses of stars need to try to unravel these uncertainties. We shall look at how they do this in the next section.

The orbit of a binary

In this section: the orbits of stars in a binary system are ellipses, just like planetary orbits.

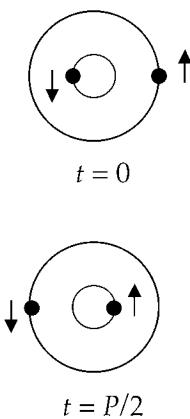


Figure 13.2. For the discussion in Investigation 13.1 on page 156, two stars in circular binary orbits shown at times half an orbital period apart.

We saw in Chapter 4 that the orbit of a planet around the Sun was a perfect ellipse, at least if the gravitational effects of other planets are ignored. It would be natural to expect that the situation would be more complicated when the "planet" is actually another star, whose mass is comparable to that of the star it is orbiting. After all, in the Solar System we idealized the Sun as being fixed at one point, undisturbed by the weak gravitational pull of the planets. In a binary star system, neither star will stand still, and so we might expect the orbits to be much more complex.

Remarkably, this is not the case.

Provided the stars in a binary system are themselves spherical, *both* of their orbits will be ellipses, just as for planets in the Solar System.

We show this for the special case of circular orbits in Investigation 13.1 on page 156. For the general elliptical case, we turn to the computer program for orbits that we constructed in Chapter 4. A simple modification in Investigation 13.2 on page 157 is enough to demonstrate the elliptical nature of the orbits of both stars. The results of two computer runs are shown in Figure 13.3.

The restriction to stars that are spherical is the same as we had in the Solar System: the gravitational field of a spherical star is the same as if all its mass were concentrated at its center. This restriction is not always valid. When the stars are close to one another, the tidal forces of one deform the shape of the other, and the orbit can become much more complex.

Computer Simulations of Binary Star Orbits

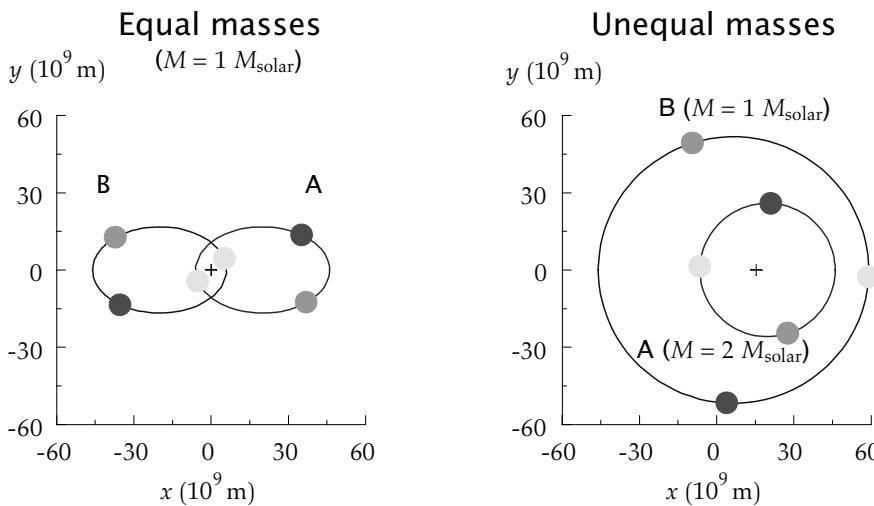


Figure 13.3. Results of computer simulations of the orbits of two binary systems. On the left is a system consisting of two $1 M_{\odot}$ stars with initial speeds of 13 km s^{-1} at their point of furthest separation, which is taken as twice the minimum separation of Mercury from the Sun. This should be compared to Mercury's speed of 59 km s^{-1} at that point. The stars plunge together rapidly. The right-hand figure shows a system in which star A has mass $2 M_{\odot}$ and B has $1 M_{\odot}$. The initial speed of A is 20 km s^{-1} , while that of B is 40 km s^{-1} . In both diagrams, the positions of the stars at selected times are illustrated by dots on the orbits. The stars' positions always lie diametrically opposite each other on a line through the common focus of the orbital ellipses, which is shown as a "+".

Planetary perturbations

To a good approximation, the Solar System can be regarded as a binary system involving the Sun and Jupiter. In this time-dependent gravitational field, the other planets make their orbits around the Sun. We have mentioned before that, although the orbits would be elliptical if the Sun were the only body creating the gravitational field, the orbits are not exactly elliptical when we take account of Jupiter's influence. To see this, I have again modified the computer program that was used to calculate Mercury's orbit in Chapter 4; the new program can follow Mercury in the gravitational field of the Sun and one other body. I assume the Sun and "Jupiter" follow circular binary orbits. The program *MercPert* is available on the website.

The effect of Jupiter on Mercury in our Solar System is rather small, although easily measurable if one uses observations of the orbit performed over a century or so. However, when the Solar System first formed, things may have been much more interesting. In particular, astronomers speculate that several Jupiter-sized planets may have formed and spiralled into the Sun rather quickly. In order to illustrate what might have happened to an inner planet in such a situation, on a short time-scale, I have re-shaped the present Solar System: it is nice how computers allow one to play with such ideas! This hypothetical planetary system has a planet 100 times as massive as Jupiter, lying in the orbit of Venus. Such a massive planet so near to Mercury makes the changes in Mercury's orbit easier to see, but similar things would happen if Jupiter were less massive. Astronomers have recently discovered planets around other stars that are not very different from this super-Jupiter.

The results of our simulation are illustrated in Figure 13.4 on page 158. In the left panel are shown the orbits of all three bodies. Notice the extraordinary event where Mercury actually moves out and loops around the planet. This was not something I set out to achieve: it happened on the first run with this configuration.

Mercury in effect indulges itself in the gravitational slingshot here. Its orbit after the encounter goes much closer to the Sun than before, just as we noted would happen to an artificial space probe that meets Jupiter in this way (Chapter 6).

In this section: each planet is a kind of binary body with the Sun, but the planets also affect each other. We construct imaginary Solar Systems and use computer simulations to see what might have happened in the early Solar System.

Investigation 13.1. Two stars making circles around each other

It frequently happens that binary orbits are nearly circular, and in that case it is not hard to find out how to use observations to give us information about the masses of the stars.

First, we need to convince ourselves that circular orbits are possible. In Figure 13.2 on page 154 I illustrate the geometry. Two stars of different masses move on circles of different radii but with a common center. The orbits have the same period, so the stars are always on opposite sides of the center. Can we insure that the system can be set up to behave like this?

Suppose we first choose an arbitrary value for their common orbital period P . We next have to decide on the sizes of their orbits. Once we choose the orbital radii, we then know what speed we have to give to the stars, which has to be just the right amount to get the stars around their orbits in the time P . The radii must be chosen so that the acceleration required to keep each star on its circular orbit equals the gravitational acceleration produced by the other star at the distance separating them (which is the sum of the orbital radii). Then each star will at least initially tend to move on its circle, and the stars will continue to lie diametrically opposite each other.

Since the speed is constant in circular motion, a short time later exactly the same conditions will continue to hold, and so the stars will continue to follow their circles. Therefore, circular motion of the type shown in Figure 13.2 on page 154 is possible, given the right initial conditions. In fact, even when the initial conditions are not right, friction in the system (such as results from tidal deformation of the stars) can circularize the orbit.

Suppose the stars, called “1” and “2”, orbit on circles of radii R_1 and R_2 , respectively. Let their masses be M_1 and M_2 . Suppose their orbital period is P , the same for both stars. Let their total separation be called $R = R_1 + R_2$. Consider the acceleration of star 1. Traveling a circle, it has uniform acceleration towards the center of $4\pi^2 R_1/P^2$, from Chapter 4. The gravitational acceleration produced by the other star is GM_2/R^2 , depending on the mass M_2 of the other star. Circular motion requires these to be equal:

$$\frac{GM_2}{R^2} = \left(\frac{2\pi}{P}\right)^2 R_1. \quad (13.1)$$

There is an analogous equation for the second star, obtained by exchanging the indices 1 and 2:

$$\frac{GM_1}{R^2} = \left(\frac{2\pi}{P}\right)^2 R_2. \quad (13.2)$$

Let us first divide these equations. That is, we divide the left-hand side of Equation 13.1 by the left-hand side of Equation 13.2 and similarly for the right-hand sides. The ratios remain equal, and most of the factors cancel out. We are left with the simple expression

$$\frac{M_2}{M_1} = \frac{R_1}{R_2}. \quad (13.3)$$

The sizes of the orbits are in inverse proportion to the masses of the stars. The heavier star executes the smaller orbit. We already knew this: we have used it in Chapter 4 to determine the radius of the circle the Sun moves on due to Jupiter’s gravitational pull.

Next we add the two equations, by adding the left-hand sides to each other and the right-hand sides to each other. This gives

$$\frac{G(M_1 + M_2)}{R^2} = \frac{4\pi^2}{P^2} R. \quad (13.4)$$

This is interesting because it shows that simply measuring the period P of the binary gives us the ratio $(M_1 + M_2)/R^3$.

If we knew the stars’ separation R we could then infer the total mass of the two stars. This is observable in a visual binary, but not in a spectroscopic binary. In fact, in a visual binary one can determine everything, since one can measure both orbital radii: these immediately give M_1 and M_2 from the original equations.

But visual binaries are not common. Normally astronomers observe a spectroscopic binary with only one set of lines, the other star being too dim to appear in the spectrum. Let us call this star 1. By replacing R in Equation 13.4 by $R_1 + R_2 = R_1(1 + R_2/R_1) = R_1(1 + M_1/M_2)$, we obtain the following useful equation:

$$\frac{M_2^3}{(M_1 + M_2)^2} = \frac{4\pi^2 R_1^3}{GP^2}. \quad (13.5)$$

Astronomers can measure the period P , and they can *almost* measure R_1 . What they actually measure is the Doppler shift of the spectral line, as in Figure 13.1 on page 154, which tells them the speed of the star along the line-of-sight. Combined with the measured period P , this tells the observer how much distance the star moves toward and away from the Earth during its orbit. But it does not reveal how much the star moves in a plane perpendicular to the line-of-sight, which astronomers call the **plane of the sky**. So they do not measure R_1 , but rather the projection of R_1 onto the line between the Earth and the star. It is easy to see that this differs from the true radius by the sine of the angle between the line-of-sight and a line perpendicular to the true plane of the orbit. Astronomers call this angle the **angle of inclination** of the orbit, and use the symbol i for it. Thus, observers measure $R_1 \sin i$. Multiplying Equation 13.5 by $\sin^3 i$ gives a right-hand side that is measurable, and therefore the left-hand side is a known quantity for spectroscopic binaries. This is called the **mass function** of the circular orbit:

$$f(M) = \frac{M_2^3 \sin^3 i}{(M_1 + M_2)^2}. \quad (13.6)$$

It has dimensions of mass, but its value only constrains the masses of the two stars. From observations of a single star in a spectroscopic binary, one cannot determine the individual masses without additional information.

Additional information is sometimes available. If the second star’s spectrum is also visible, then its velocity determines the mass function with indices “1” and “2” reversed; this allows the mass ratio of the two stars to be determined. In a single-spectrum system, the visible star may have a standard and well-understood spectrum, so that theoretical calculations determine its mass; then the mass function can be used to constrain the value of the companion’s mass. Alternatively, the companion star may pass right in front of the star that we see, eclipsing it; this requires that the angle i is nearly 90° .

We will see another example of extra information in Chapter 20, when we consider binaries containing pulsars. For very compact binaries, general relativity predicts extra effects that can be measured and used to determine the individual masses and the angle i .

In the right-hand panel of the figure, I show the orbit of Mercury relative to the Sun; that is, I plot the x - and y -distance of Mercury from the Sun. This is not identical to the path of Mercury in the left-hand panel, because the Sun moves. After the near encounter with the planet, Mercury’s orbit is considerably more elliptical than it would have been with no planet there. It also does not keep its orientation in space: the (imperfect) “ellipse” it traces out turns counterclockwise. The likelihood exists that a further encounter with the massive planet would send Mercury plunging into the Sun.

Interested readers are encouraged to play with this program. There is an infinite

Investigation 13.2. Simulating the orbits of a binary pair

Binary orbits are generally not circular. To demonstrate to ourselves that they are actually elliptical, we do the same as we did before: we simulate their orbits using the computer, and then look at the shape of the orbits. This cannot be exact, but it can be made very convincing.

On the website is the program *Binary* that does this calculation. It is adapted from the program *Orbit*, which did the orbit of Mercury that was displayed in Figure 4.3 on page 29. The adaptation is straightforward. The main complication is that there are now two bodies to follow. I have called them A and B, and their coordinates, for example, are (x_A, y_A) and (x_B, y_B) . Each statement in *Orbit* that refers to the coordinates, velocity, or acceleration of the body has become two statements, one referring to body A and the other to body B. Tests for, say, the appropriate time-step size are performed on both orbits, and both must pass the test.

The main change in the physics underlying this program is the fact that the gravitational acceleration no longer comes from the Sun, located at a fixed point which we took to be the origin of the coordinate system. Instead, the attraction on body A comes from body B, and vice versa. We must therefore insure that, no matter how the bodies move, the acceleration of gravity is calculated correctly.

This change is not hard to make. Let us write down again the equations we used for the acceleration produced by the Sun, Equations 4.5 and 4.6 on page 29:

$$\begin{aligned} x\text{-accel} &= -k \frac{x}{r^3}, \\ y\text{-accel} &= -k \frac{y}{r^3}. \end{aligned}$$

What do the terms in this equation mean? The denominator r is just the distance from the planet to the Sun. In the binary problem, we would replace this with r_{AB} , the distance between the two

stars. The factors of x and y in the numerators of the planetary acceleration are the coordinates of the position of the planet. Now, the coordinates themselves have no absolute significance: they only show where the planet is in relation to the Sun, which in the earlier calculation was at the origin of the coordinate system. If the Sun, or another body exerting a gravitational force, were at a position with coordinates (x_S, y_S) , then we would replace x by the x -distance from the Sun to the planet, which is $x - x_S$. Similarly we would use $y - y_S$ in place of y .

In the binary problem, the acceleration of body A produced by B is therefore found by replacing x by $x_B - x_A$ and y by $y_B - y_A$. Similarly, the acceleration of B produced by A is found by replacing x by $x_A - x_B$ and y by $y_A - y_B$. This is done in the program.

The output of the simulation, as displayed in Figure 13.3 on page 155, shows the elliptical nature of both orbits clearly, even when the stars have different masses. Notice that the ellipses share one focus, indicated by the “+” in the diagrams. The stars remain on opposite sides of this focus at all times. In our study of the circular orbit problem in Investigation 13.1, we saw that the bodies remained on opposite sides of the origin; this is the focus of the circular orbits. The distances of the stars from the center of the circle were inversely proportional to their masses: $M_1 r_1 = M_2 r_2$. The same is exactly true for the elliptical orbits: the distances of the stars at any time from the focus are in inverse proportion to their masses.

I must not omit one small but important point about using the binary orbit program. When choosing initial speeds for the stars, make sure that the total *momentum* vanishes:

$$M_A U_A + M_B U_B = 0, \quad M_A V_A + M_B V_B = 0. \quad (13.7)$$

If initial data are chosen that violate this, then the stars will still orbit one another, but the whole system will move as well! We will see an example of a moving binary in a later analysis.

variety of possible configurations to try, and interesting things will happen very often. In particular, try replacing Mercury with a comet in a very elongated orbit plunging toward the Sun. If it comes near the planet, it could encounter it and end up in a much more *circular* orbit.

We have met here a rather complicated situation, where there are three bodies in mutual gravitational fields. In fact, this is far beyond the reach of pen-and-paper mathematical calculations, and it would not normally be treated in, say, courses leading to an undergraduate astronomy degree. But, given the original orbit program, it is hardly any extra work to get our computers to show us what we might expect to happen here. Our computers open up whole new subjects for us to think about.

Of course, the complexity of the problem does still have an impact: one really has to sit down with the computer and run many versions with different initial data in order to develop a feeling for the variety of things that might happen. I encourage the reader to do this, since the variety is really very large and interesting.

Tidal forces in binary systems

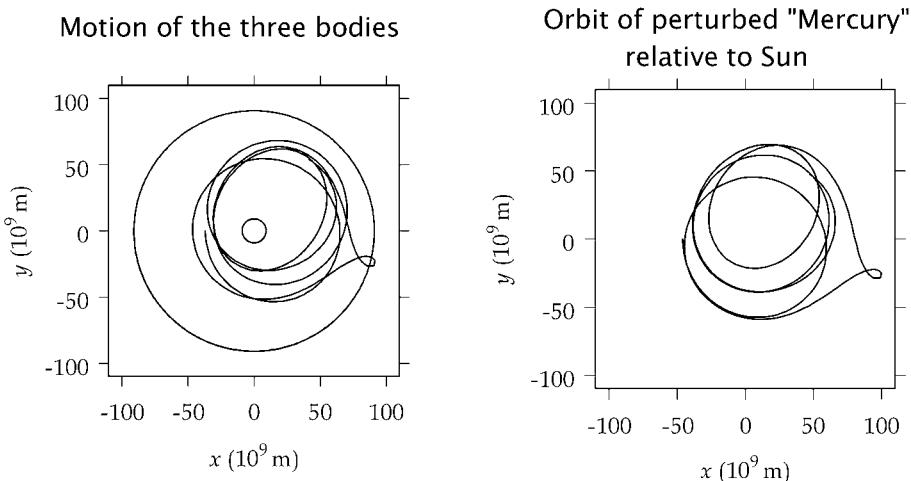
The effect of “Jupiter” on Mercury is essentially the action of the tidal gravitational forces exerted by the large planet on the orbit of the smaller body. There are many other places in astronomy where such tidal forces are important. We met some in Chapter 5. Here are some more.

>Our simulated comet will not, however, be captured by the planet, as happened to comet Shoemaker-Levy 9, causing it to crash spectacularly into Jupiter in 1994. The reason is that capturing a comet requires that some of the comet's energy be given up to friction, as when a comet breaks apart. In our computer program, there is nothing to simulate friction.

In this section: tidal forces can make gas flow from one star to another, emitting X-rays. They can help one star to capture another into a binary. And they explain why the inner planets like the Earth became rocky.

- *Mass transfer between binary stars.* One of the most dramatic examples of tidal forces occurs in binary star systems, where one star begins to pull mass off the other. We have idealized our binary systems above by assuming the stars are spherical. But stars are not rigid, so if they get too close, the tidal

Figure 13.4. Computer simulation of the effect that an imaginary massive planet in Venus' orbit would have on Mercury. All bodies orbit counterclockwise. Mercury begins between the Sun (which executes the small circular orbit) and the planet (large circular orbit), and on its second orbit happens to interact strongly with the planet, looping around it.



forces of each star will deform the other. If one star evolves into a giant, the result can be that the tidal forces actually pull the weakly bound outer layers off the giant and onto the other star. The numerical simulation in Figure 13.4 shows how this can happen. Suppose that the body we have called "Mercury" is really just a bit of the outer parts of a giant star sitting where the Sun is. Then the companion star pulls this bit of the giant over to itself (Mercury loops around "Jupiter"). What will happen next is different for gas than for a planet like Mercury. Instead of just going once around "Jupiter" and heading back to the original star, as Mercury does, a parcel of gas will run into other gas that has been pulled off before. It will stop and sink into a disk around the companion. This is called an **accretion disk**. We will look at two examples of such systems in more detail below: **cataclysmic variable** stars and **X-ray binary** stars.

- *Tidal capture.* Suppose two stars that are not originally in orbit about one another happen to pass very close to each other. If the stars were point particles, they would deflect each other's direction of motion, but then they would separate again and fly off on their own trajectories. But real stars will affect each other tidally. The tidal forces may pull material off one or both stars; they will stretch each star and cause frictional heating; they could also change the intrinsic rotational motion of one another. All of these effects can remove energy from the stars' motion, with the result that they may not have quite enough to get away from each other afterwards. They then fall into a highly eccentric elliptical orbit, repeatedly passing as close to each other as before. On each pass, they lose a bit more energy, so the orbit becomes smaller and more circular. This is called tidal capture, and it is thought to happen to a significant number of stars in the centers of the **globular clusters** of stars that we will study in Chapter 14. It also operates in the star-forming simulations shown in Figure 12.2 on page 138, where long chains of protostars merge into one another.
- *Why the inner planets are rocky.* Although the formation of the planets is still shrouded in some uncertainty, one thing is clear: they did not simply form as big balls of diffuse gas that subsequently evaporated away and left rocky cores. Instead, they probably formed by the agglomeration of small

Investigation 13.3. Forming rocky planets

We saw in Investigation 5.1 on page 43 that the tidal force exerted by a body of mass M on another body at a distance r from it is proportional to M/r^3 . This in turn is proportional to the *mean density* of the body, if its mass were spread over the entire region out to r .

From this it is easy to see what would happen to a cloud of gas condensing near the Sun. Consider the tidal force of the Sun on the outer part of the condensation, relative to the gravitational force of the condensation itself. This self-force is nothing more than the tidal force of the body on itself, so the two forces are in proportion to the two relevant densities. The tidal force of the Sun is proportional to the mean density of the Sun spread out over the whole region inside the orbit of the condensation. The force of the condensation on itself is proportional to its own density.

Now, in the original collapsing gas cloud from which the Sun and planets formed, the density must have been higher near the center

than near the edge. That means that the mean density of the gas inside the orbit of the condensation must have been larger than the density at the condensation. The gas inside went on to form the Sun, but its tidal effect on the condensation didn't change. Since it had the larger density, its tidal effect dominated. The condensation would therefore have had a very hard time forming.

If the protoplanet formed from rocky condensations that stuck to one another when they collided, then the argument would be different. The density of the asteroids was much higher than the density of the gas as a whole, and they would not have been torn apart by the Sun. Whatever the details were, it is clear that gravitational tides would have stopped purely gravitational condensations. Some chemical processes in the early nebula were required to produce the asteroids, which could then form seeds for the planets.

rocky asteroids that formed from the interstellar dust grains in the cloud of gas that formed the Sun. The reason is the Sun's tidal effect on the disk in which the planets formed. The Sun would always have been strong enough to have torn apart any ball of gas that was condensing into a planet. This is explained in Investigation 13.3. For the inner planets (out to Mars), this meant that they grew with little gas present; their present atmospheres come from the release of gas that was trapped in the original rocky asteroids. The more gaseous outer planets managed to trap some of the gas of the original cloud, but only after they had developed massive cores from the asteroids. This was easier at a great distance from the Sun, possibly because the gas there was colder. The present-day asteroid belt is probably a planet that never formed: the combined tidal forces of the Sun and Jupiter prevented even the rocky asteroids from accumulating into a planet.

Accretion disks in binaries

Once accretion disks have formed around certain kinds of stars, the subsequent events can be dramatic. Material that falls into a disk quickly gets pushed into a circular orbit around the central star, since that minimizes friction with the material already there. But since circular orbits at different distances from the central star have different periods, there is inevitably some friction always present. The result is that the material slowly spirals into the center, and its orbital energy is converted into heat. Accretion disks around compact stars, such as white dwarfs and neutron stars, are very hot near the central star.

We show in Investigation 13.4 on page 161 that the random thermal kinetic energy of particles in the disk is determined by a balance between the rate at which mass moving through the disk releases gravitational energy and the rate at which that energy can be radiated away from the disk. We find that the temperature near the center of a disk around a white dwarf of mass $1M_{\odot}$ and radius 5×10^6 m that is accreting $10^{-10} M_{\odot}$ per year is 8×10^4 K. The typical thermal kinetic energy of the particles is about 7 eV. Such a disk would emit thermal radiation in the near-ultraviolet part of the spectrum, at a wavelength of $0.17 \mu\text{m}$.

Accretion disks around white dwarfs are responsible for a wide range of observed phenomena. The material that falls on the dwarf from a normal companion comes from the companion's atmosphere, so it is mostly hydrogen. When it lands on the white dwarf, it is hot and much denser. After a certain amount has accumulated on the surface of the dwarf, a nuclear chain reaction can occur, converting the hydrogen to helium. This sudden release of energy causes an explosion, which we see as a nova. This is very different from a supernova, in which a whole star is disrupted. In

In this section: an accretion disk can emit X-rays, blow up as a nova, or funnel mass onto a central star until it becomes a supernova.

a nova, the disruption is temporary; accretion soon resumes, and it is only a matter of time until the next nova occurs in the same system.

Accretion disks can be responsible for supernovae, as well. If the accretion rate is high enough, then after a long period of accretion and despite many nova outbursts, the white dwarf's mass will have increased to the Chandrasekhar mass, and the star will collapse. As we saw in the previous chapter, this white dwarf typically contains much material that can still undergo nuclear reactions, and the result can be a giant nuclear reaction that incinerates the whole star. This is believed to produce Type Ia supernova explosions.

Between nova explosions, an accreting white dwarf can still appear to be unusual. The flow of material onto the star can be irregular for many reasons, and minor outbursts can occur with surprising regularity. Such binary systems are called *cataclysmic variables*.

Compact-object binaries

In this section: the most spectacular accretion phenomena occur when the central star is a neutron star or black hole.

Observing the emitted X-rays is one of the main ways of identifying black holes.

If we took the same accretion disk as we placed around a white dwarf in the previous section and put it around a neutron star of radius 10 km, it would have a central temperature of 8×10^6 K, and would radiate thermal X-rays. The typical thermal energy of a particle would be 0.7 keV. (The notation "keV" means kiloelectron volts, or 10^3 eV.) This is called a "soft" X-ray. We see such systems because they are strong emitters of X-rays. There are over 100 so-called X-ray binaries in our Milky Way galaxy.

If one can identify the "normal" star (usually a giant) that provides the gas for the accretion disk, then one can usually measure the Doppler shift of its spectrum as it orbits the neutron star. This gives us some information about the mass of the neutron star, as we have explained earlier. It is quite remarkable that all neutron stars identified in this way have masses of about $1.4M_{\odot}$.

The compact object at the center of an accretion disk could also be a black hole. Here the energy would be a little higher, but the system would still be an X-ray source. Since black holes are likely to be formed when the collapse to a neutron star involves more mass than the maximum for neutron stars, black holes are expected to be more massive than neutron stars. We now observe a number of systems where the mass of the compact object is likely to be $8M_{\odot}$ or more. Since this is much higher than the maximum mass of neutron stars, these are identified as black holes. We shall describe black holes in Chapter 21.

Fun with the three-body problem

In this section: we modify the computer program so it can simulate three stars of similar masses interacting with one another. Spectacular consequences follow. We simulate a "factory" for black hole binary systems.

We have studied above the problem of three bodies interacting by gravity, which astronomers call the "three-body problem". However, we have not made it the most general three-body problem we might imagine, because we have not let the gravitational field of Mercury affect the other two bodies. We have treated Mercury as if it were a "test particle", a probe of the field created by the other two bodies. Astronomers therefore call this the "restricted three-body problem".

This is all right if Mercury's mass is small, as it is here. But the general case is even more interesting, and we turn to it now. The full three-body problem opens up a new possibility that the restricted problem does not have. Although a system of three stars may start out mutually bound, in the sense that none of them has the escape velocity to get away from the other two, it is possible for two of them to form a very close binary pair, giving the third such a strong "kick" that it leaves the system altogether. In fact, such behavior is the norm rather than the exception.

I have modified the orbit program to calculate the full three-body problem. This means treating each star the same, and allowing its gravitational forces to act on

Investigation 13.4. Accretion disks

We have learned enough physics in previous chapters to predict fairly accurately the temperature and observable features of accretion disks. The essential feature of an accretion disk is that material flows through it, gradually falling onto the central star. This flow is driven by friction. The mechanism creating the friction could be complicated, but it seems to be the case that many accretion disks are in a fairly steady state, so that the mass flows through the disk at a constant rate.

What makes accretion disks hot? They are hot, of course, because gravitational energy is being released by the material falling in. Now, a disk without friction would consist just of circularly orbiting gas: no matter would flow through it onto the central star, no energy would be released, and the disk would have zero temperature. (Think of the rings of Saturn here.) An accretion disk can remain hot, with a steady luminosity, only if material flows through it. The reason is that it must constantly replace the energy it radiates away with new gravitational energy released by the flow of matter through it. The flowing gas must, of course, come from somewhere and go to somewhere. The gas that reaches the inner edge of the disk falls onto the central star or into the central black hole. If the disk is in a steady state, then this gas has to be replaced by new gas arriving at the outer edge of the disk. Normally this gas comes from a companion star in a binary system.

Let us look at what happens in a steady disk. Suppose that, in a small interval of time Δt , an amount of mass ΔM arrives at the outer edge of the disk, and the same amount leaves from the inner edge. For a steady situation, the mass arriving, ΔM , will be proportional to the time interval Δt . If we call the constant of proportionality M_t , then we can write^a

$$\Delta M = M_t \Delta t.$$

Now, each small amount of mass m that reaches the inner edge of the accretion disk, whose radius is R , has had to release a total energy roughly equal to the energy it would need to escape from this radius, because escaping is just the time-reverse of falling in. (I say roughly, not exactly, because the gas still has orbital kinetic energy and maybe a little inward speed as well at the inner edge. If it encounters the central star there, then all this energy will be released as it hits the star. But if the central object is a black hole, then the kinetic energy will go into the hole and not be converted into radiation. This would roughly halve the energy released.) This escape energy is just the gravitational potential energy at the inner edge $E = GMm/R$. Then if ΔM amount of mass moves through the disk in the time Δt , it releases an energy $\Delta E = GM\Delta M/R$. This energy is released all over the disk, but more is released near the inner edge, where the gravitational potential energy is largest.

Now, the *luminosity* L of the disk equals the energy released per unit time, i.e. $L = \Delta E/\Delta t$. Using the fact that $\Delta M/\Delta t = M_t$ allows us to write

$$L = GMM_t/R. \quad (13.8)$$

To find the disk's *temperature*, we assume that the disk is a *black body*. Knowing the luminosity of the black body, we need to estimate its surface area in order to deduce its temperature. Although the accreting matter will release its energy over the whole of the disk as it gradually spirals through it, most of the energy is released in

the central region, especially if the material finally accretes onto the surface of a central star. We won't be far wrong, therefore, if we assume that the area is πR^2 . This is not exact, but it will give us a good idea of what temperatures to expect.

Equating the energy released to the energy radiated at temperature T gives

$$\frac{GMM_t}{R} = \sigma(\pi R^2)T^4. \quad (13.9)$$

Solving this for the temperature gives

$$T = \left(\frac{GMM_t}{\pi\sigma R^3} \right)^{1/4}. \quad (13.10)$$

This equation leads to the numbers quoted in the text.

Astronomers often use formulas like Equation 13.10 above that can be applied over a wide range of values of some of the parameters. For example, one astronomer might be dealing with accretion onto a white dwarf, another with accretion onto a neutron star; the difference in R is a factor of 600. Similarly, different sorts of binary systems could have values of M_t that differ by factors of 1000, depending on the mass and size of the companion star. It is useful in such situations to pick values of the parameters suited to one situation, calculate the desired numbers, and then show how the result would scale with changes in the values of the parameters. In this way, we do the arithmetic involving numbers that don't change (like G) once and for all. For example, if we take $M = 1 M_\odot$, $M_t = 10^{-10} M_\odot \text{ yr}^{-1}$, and $R = 5 \times 10^6 \text{ km}$, then we get $T = 8 \times 10^4 \text{ K}$. Having calculated this once, there is no need to go through the trouble of looking up values of G and σ again for a different central stellar radius. Instead, we express the general result in the following way:

$$T = 8 \times 10^4 \left(\frac{M}{1 M_\odot} \right)^{1/4} \left(\frac{M_t}{10^{-10} M_\odot \text{ yr}^{-1}} \right)^{1/4} \times \left(\frac{R}{5 \times 10^6 \text{ m}} \right)^{-3/4}. \quad (13.11)$$

The proof that this is equivalent to Equation 13.10 is that (1) when the values assumed above are used, the answer is right; and (2) the value of T in Equation 13.11 depends on the variables M , M_t , and R in the same way as in Equation 13.10.

Now when we change the parameters, we can do the calculation more easily. For example, take the same accretion rate onto a solar-mass neutron star with $R = 10 \text{ km}$. All values are the same except for R , which is a factor of 500 smaller. Since the temperature depends on $R^{3/4}$, we conclude immediately that the temperature is a factor of $500^{3/4} = 105$ larger. Recall that the wavelength of the emitted radiation depends inversely on the temperature. A temperature of just under 10^5 K , as in Equation 13.11, is cooler than the Sun and therefore emits predominantly infrared radiation. When we replace the white dwarf by a neutron star, the wavelength goes right down into the X-ray region of the spectrum. So when astronomers first began observing X-rays, they began to find neutron stars and black holes in binary systems, not white dwarfs.

Exercise 13.4.1: Accretion disk temperatures

Perform the arithmetic to get the temperature $T = 8 \times 10^4 \text{ K}$ in Equation 13.11. Then use this equation to calculate the other values of temperature given in the text.

Exercise 13.4.2: Accretion disk luminosities

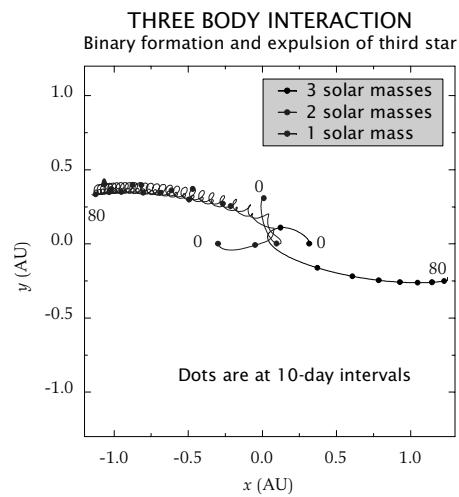
Take the equation $L = GMM_t/R$ for the disk luminosity and write it in a similar normalized form to Equation 13.11, scaling the mass of the central object to $10^9 M_\odot$, the accretion rate M_t to $1 M_\odot \text{ yr}^{-1}$, and the radius to 10^{13} m . These values are appropriate to accretion disks around the giant black holes that power quasars (Chapter 14).

^aReaders who are familiar with calculus will recognize M_t as the derivative of M with respect to t , dM/dt , which is constant for a steady flow. Other readers will not need to know this.

every other star. The program `Multiple`, available on the website, is written so that any number of stars can be used, so it can investigate the four-body, five-body, and in general the n -body problem. Of course, the more bodies that one uses, the slower the program runs: since each body's motion must take account of the forces from every other body, the number of calculations necessary to advance one time-step is roughly proportional to n^2 for n bodies. So the four-body problem takes nearly twice as long to run as the three-body problem. I would encourage only those readers with fast computers to go beyond three bodies.

The result of the first run I made with the three-body program is shown in Figure 13.5. I took three stars of masses 1, 2, and $3M_{\odot}$, and arranged them with separations comparable to the Mercury–Sun distance. I gave them small initial velocities, to insure the system was bound overall: no star had the escape speed from the gravitational field of the other two. All motion was in a single plane. The result, as is evident from the chart, is that the two smaller stars formed a bound binary pair, and expelled the third star. The pair runs off to the left in orbit around each other, and the single star moves the other way. The system is no longer bound: it breaks into two pieces.

Figure 13.5. Expulsion of a star from a three-body system. The result of the simulation described in the text is that two stars form a close binary and expel the third. The starting positions are marked “0”, and their positions after 80 days are marked “80”. The axes are calibrated in astronomical units (AU).



the pair: the result in this case is to bind the pair more tightly and expel the third. Under other circumstances, the result could have been a different binary pair, but the remaining single star would still have been less tightly bound, and may have been expelled. I encourage the reader to experiment with different initial conditions, just to get a feeling for the frequency with which stars are expelled from triple systems.

Such events do happen in astronomical systems. We shall meet globular clusters – dense systems of millions of stars – in the next chapter. Black holes tend to settle into their centers, where they occasionally – over millions of years – undergo close three-body encounters. Astronomers speculate that such collisions could lead to the formation of numerous close binary black holes, and in the first decade of the twenty-first century astronomers will be searching for tell-tale gravitational waves from such systems. In the next chapter we will look at the globular clusters that produce these black holes, and at their galaxies, which contain even more massive central black holes that might have formed in similar ways; in Chapter 22 we will learn what these waves of gravity really are.

How can it happen that a system that is initially bound later becomes unbound? The answer is in gravitational energy. When two stars become more tightly bound, they release energy, just as a particle falling onto a star releases energy. This energy must go somewhere, and it goes into the motion of the third star. This is not by some magic. If the third star were not present, the first two could not form a tight binary pair: they would fall towards one another and then recede to the same distance. The forces exerted by the third star allow the outcome to be different: it acts as the “marriage broker”. By Newton's third law, the force exerted on the pair by the third star is exerted back on the third star by

Galaxies: atoms in the Universe

We are now ready to make another step outwards in our exploration of the Universe: we change from looking at stars to looking at *galaxies*. As we saw in Chapter 9, galaxies are vast collections of stars. Our own Galaxy, the familiar Milky Way, contains about 10^{11} stars. Figure 14.1 on the following page shows photographs of two typical galaxies. Galaxies are held together by the mutual gravitational attraction of all the stars. It is remarkable indeed that Newton's force of gravity, which he devised in order to explain what held the planets in their orbits, turns out to explain just as well what holds the whole Milky Way together.

Galaxies are more than just collections of stars. The collective gravity of all the stars makes the centers of galaxies very unusual places. Stars and gas crowd together so densely that in some cases they can form immense black holes, with masses of millions or even billions of stars. These black holes then become the sites of intense activity: as gas falls towards them and heats up, it can shine more brightly than the whole remainder of the galaxy. Even more astonishing are the **jets**: two collimated streams of ionized gas, shooting in opposite directions out of the centers of some galaxies at nearly the speed of light, maintaining their intensity and direction for millions of years. We shall study below various ways that this activity shows itself: quasars, giant radio galaxies, Seyfert galaxies.

Galaxies also hide enormous amounts of dark matter. We don't yet know what this dark matter might be, but galaxies provide us with the evidence that it is there: it produces much more gravity than can be explained by the stars that we can see. The problem of the **missing mass** is one of the most intriguing in all of astrophysics.

Galaxies are, of course, very large. When we jumped from the Solar System to the nearest star, we increased the scale of distances by a factor of some 10^5 . In this chapter we jump by an even larger factor. The size of a galaxy is already a factor of 10^4 larger than our previous distance-scale: the typical size of a galaxy is tens of kiloparsecs (kpc), compared to the typical distances between stars of 1 pc. And the separations between galaxies are larger by a further factor of 100, up to the megaparsec (Mpc) scale. Despite these enormous distances, the concentrating force of gravity in the centers of galaxies often leads to intense activity that takes place in regions the size of our Solar System.

Galaxies mean another kind of jump for us: a jump in time to modern astronomy. In earlier chapters, when we described the Solar System, we were on ground that would have been familiar to Newton and Galileo. When we studied stars, we dealt with issues that would not have surprised most nineteenth century astronomers. Even the first speculations about black holes and the gravitational deflection of light belong to the eighteenth century. Modern astronomy and astrophysics have made huge advances in our *understanding* of the planets and stars, but the objects themselves were part of the known Universe of the nineteenth century.

In this chapter: we finally reach the basic building blocks of the Universe: galaxies. Galaxies come in many shapes and sizes. They foster the formation of stars and harbor giant black holes in their centers. They contain only some of the mass in the Universe: much more is dark and unidentified. As beacons of light, they allow astronomers to measure how rapidly the Universe is expanding. Their first stages of formation are imprinted on the cosmic microwave background radiation.

►The image under the text on this page shows our Milky Way galaxy as seen in the radio waves called microwaves. It was compiled from observations by the COBE satellite. The view shows the full 360° sky around our location, flattened onto a single view. Courtesy NASA Goddard.

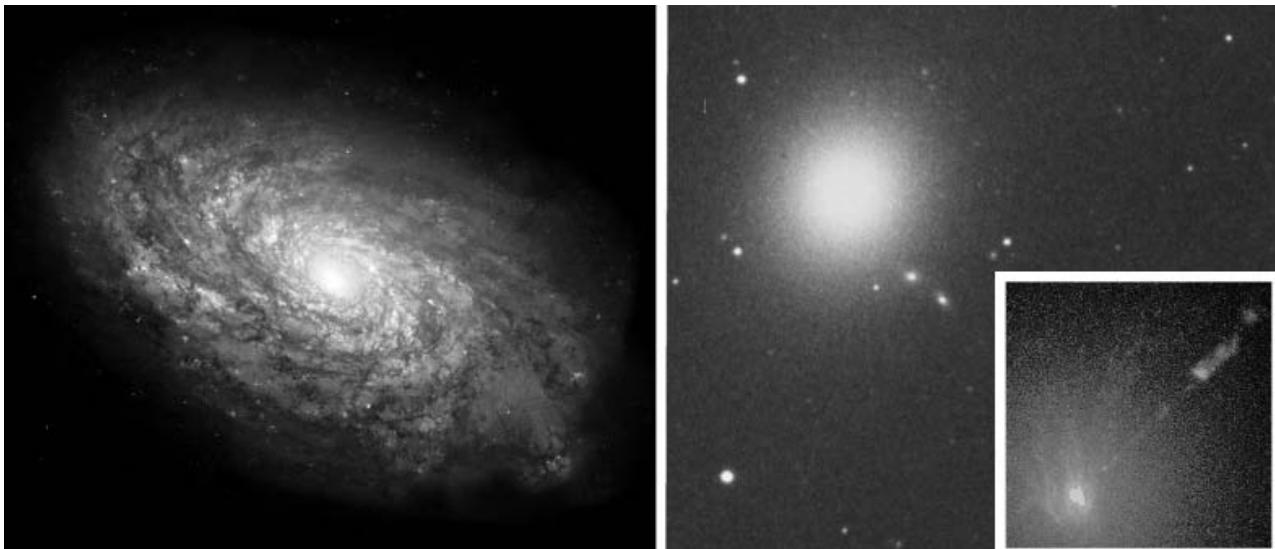


Figure 14.1. Photographs of two galaxies that are representative of the two main types seen by astronomers. The galaxy on the left is a spiral galaxy known as NGC4414. It is very similar in form to the Milky Way. The galaxy on the right is an elliptical galaxy known as M87. It is one of the largest galaxies in the Virgo Cluster of galaxies. It has ten times as many stars as the Milky Way. The small fuzzy objects near it are globular clusters in orbit around it. Despite its smooth outer appearance, M87's center contains a jet of gas moving outwards at close to the speed of light. This is illustrated in the inset photo. The jet indicates the presence of a massive black hole at the center. We shall see below that studies of the motion of gas near the center of the galaxy indicate that the black hole has a mass larger than $10^9 M_\odot$! All photographs courtesy NASA/STScI. The photos of NGC4414 and the jet in M87 are HST images, courtesy NASA/STScI. The photo of M87 is courtesy of the Palomar Observatory and Digitized Sky Survey created by the Space Telescope Science Institute, operated by AURA, Inc. for NASA and is reproduced here with permission from AURA/STScI.

Galaxies, on the other hand, were not recognized as being stellar systems outside the Milky Way until the twentieth century. And the discovery of their incredible activity had to await the opening of new windows on the Universe: radio astronomy, X-ray astronomy, and the use of Earth-orbiting astronomical observatories. The study of galaxies is quintessentially modern.

Globular clusters: minigalaxies within galaxies

In this section: globular clusters are small self-contained star systems in our Galaxy. They may have been building blocks from which the Milky Way was assembled. They are fragile and easily disrupted, but also seem to be rich factories of binary stars and black holes.

It is useful to begin our study of galaxies by looking at globular clusters, which share some of the properties of galaxies on a smaller scale. A representative globular cluster is illustrated in Figure 14.2. They are clusters of typically a million stars, all formed at about the same time, held together by their mutual gravitation.

Although the picture looks crowded, the distances between stars in a cluster are still very much larger than the sizes of stars, so direct collisions are rare. Distant encounters between stars can, however, still transfer small amounts of energy between them, and in globular clusters it generally takes less than a billion years for such encounters to share out the energy of the stars randomly. Scientists call this time-scale the **relaxation time**, and they say that globular clusters are **relaxed**.

This means that the velocities of stars at any point inside the cluster are fairly random, in direction as well as size. The stars form what is called a **collisionless gas**. Unlike the air in a room, where gas molecules collide very frequently, stars in globular clusters collide directly almost never. This is fortunate: gas molecules survive their collisions unharmed, but stars would be completely destroyed!

Globular clusters are prized by astronomers as museums of stars. Because they have held their “collections” intact since they were formed, they are excellent places to test ideas about stellar evolution. Hertzsprung–Russell diagrams (Chapter 12) are particularly interesting for

Investigation 14.1. Boiling away a globular cluster

We shall calculate here roughly how much energy is required to "boil away" the stars from a globular cluster. A typical cluster contains $M_{\text{cl}} = 10^6$ stars of average mass $1M_{\odot}$, in a sphere of radius $R_{\text{cl}} = 50$ pc. The escape velocity from such a cluster is

$$v_{\text{escape}} = \left(\frac{2GM_{\text{cl}}}{R_{\text{cl}}} \right)^{1/2} = 13 \text{ km s}^{-1}. \quad (14.1)$$

The energy a star of mass $m = 1M_{\odot}$ needs to escape is then

$$E_{\text{escape}} = \frac{1}{2}mv_{\text{escape}}^2 = 2 \times 10^{38} \text{ J}. \quad (14.2)$$

To boil off the whole cluster means adding roughly this energy to every star, which requires a total energy of

$$E_{\text{boil}} = 10^6 E_{\text{escape}} = 2 \times 10^{44} \text{ J}. \quad (14.3)$$

This may seem like a large amount of energy, but we must compare it to other kinds of energy that may be available, such as the energy released when stars form close binary pairs.

The energy released by a binary pair when the pair is formed is the same as the energy required to split it up again. The calculation is similar to the previous one. If the stars have an orbital separation R , then the escape velocity of star 1 (mass M_1) from star 2 (mass M_2) satisfies $v_{\text{escape}}^2 = GM_2/R$. The energy of star 1 when it has this speed is

$$E_{\text{binary}} = \frac{1}{2}M_2 v_{\text{escape}}^2 = GM_1 M_2 / 2R. \quad (14.4)$$

Exercise 14.1.1: Binding energy of a cluster

Show that E_{boil} can be expressed as

$$E_{\text{boil}} = \frac{GM_{\text{cl}}^2}{R_{\text{cl}}}. \quad (14.5)$$

In this form it is usually called the *binding energy* of the cluster. This is only an approximation, of course, accurate to a factor of two or so.

globular clusters: because all the stars have the same age, they show the relative rate of evolution of stars of different masses. The oldest globular clusters also tell us what the abundance of elements was in the gas that the first generation of stars formed from.

Despite the fact that some globular clusters have been around since the Milky Way formed, they are not robust structures. We show in Investigation 14.1 that they can in fact be disrupted completely by the energy that is released when a single close neutron star or white dwarf binary system is formed inside them, perhaps as a result of a chance three-body encounter like the one simulated in Figure 13.5 on page 162. Given their fragility, it is possible that most of the initial globular clusters of the Milky Way have already been disrupted, either by internal events as in Investigation 14.1 or by the tidal gravitational effects of other clusters or the Galaxy itself.

Globular clusters are, however, not merely museums. They process many of their stars in unusual ways. The most massive stars gradually sink towards the center, giving up energy to lighter stars by gravitational interactions. The most massive objects normally formed in globular clusters are black holes, so over a period of time the centers of globular clusters become rich in black holes. Three-body collisions among such holes (again as in Figure 13.5 on page 162) can form a binary black hole system, and some of these might be observed by gravitational wave detectors, as we shall discuss in Chapter 22.

Describing galaxies

Like virtually everything else in astronomy, galaxies come in a wide variety of shapes and sizes. Figure 14.1 illustrates the two main types, spiral galaxies and **elliptical galaxies**. Spirals have a central bulge surrounded by a wide, thin disk.

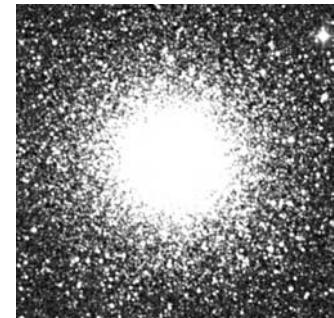


Figure 14.2. Photograph of the globular cluster M3. Photo courtesy of the Palomar Observatory and Digitized Sky Survey created by the Space Telescope Science Institute, operated by AURA, Inc. for NASA and is reproduced here with permission from AURA/STScI.

Suppose the stars each have one solar mass and are separated by the radius of a white dwarf, about 5×10^6 m. This could represent a very close white dwarf pair or a well-separated neutron star binary. Then this evaluates to 3×10^{43} J.

This is already about 10% of the binding energy of the cluster, and it is only one binary system. The formation of a handful of such systems could easily provide enough energy to expand the cluster or even disrupt it. And if a very close pair of neutron stars is formed, with a separation of, say, 100 km (still 10 times the neutron star radius), the energy released would be seven times as much as would be required to boil off all the stars from the cluster!

How would the energy released in this way get transferred to all the stars in the cluster? We saw in Figure 13.5 on page 162 that when a close pair is formed in a three-body encounter, the excess energy goes to the third one. If this happened in a cluster, the third body would have so much energy that it would simply shoot straight out of the cluster and leave the cluster essentially unchanged. But if instead the binary is first formed with a relatively wide orbit, perhaps highly eccentric, then it could shrink by a succession of much smaller energy transfers to other stars in its neighborhood. These stars would not receive enough energy to escape the cluster, and so eventually they would transfer their energy to the cluster stars generally, and the cluster as a whole would change.

In this section: most galaxies come in one of two basic shapes, spiral and elliptical. Ellipticals may result from mergers of spirals.

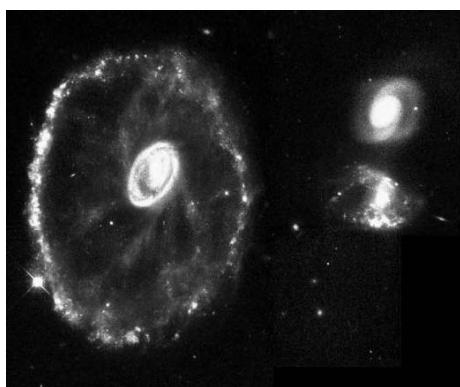
The bulge has a “hot” stellar distribution, in the sense that the random kinetic energy of a typical star is comparable to the kinetic energy it needs to orbit about the center of the galaxy. The disk is cold, since the stars have small random velocities compared to their orbital speeds. Besides its stars, the disk generally has lots of gas and dust.

Spiral galaxies take their name from the striking spiral patterns seen in many of them. It appears that the distribution of stars is actually much more symmetrical about the central axis than these patterns might suggest. The patterns instead trace out the places where new stars are being formed.

This is an interesting illustration of what astronomers call a **selection effect**. We will see below that the size and mass of a typical spiral are such that the orbital period of a star in the disk about the center of the galaxy is of the order of 10^8 years. But we saw in Chapter 12 that massive stars live only about 10^6 years. These massive stars form very bright giants at the end of their lives, which stand out much more in a photograph. An observation limited to a certain minimum brightness will inevitably *select* a much larger fraction of the bright stars than of the ordinary ones. So the spiral features are the locations of the brightest stars, not the vastly more numerous ordinary ones. And since these stars live only a short time, we are seeing them where they form.

Scientists understand some aspects of why the star-forming regions of many galaxies have such a regular spiral shape. They generally agree that this has to do with a **density wave** of some sort, which moves through the disk and compresses molecular clouds, triggering star formation. But how the wave is maintained, and what the triggering mechanism is, are still matters of debate. It should also be remarked that many galaxies do not have spiral star-forming regions: these regions can be much more irregular. **Irregular galaxies** are the third broad classification of galaxy appearance.

Figure 14.3. A head-on collision between two galaxies. The galaxy on the left was a spiral galaxy before one of the two galaxies on the right passed directly through its center, sending a shock wave outwards through the gas of the galaxy. As the wave travels out, it triggers star formation, resulting in a ring of bright young stars. The ring gives the galaxy its name: the Cartwheel Galaxy. Photograph by the HST courtesy NASA/STSCI.



Elliptical galaxies, by contrast, seem to be virtually free of gas and dust. They look more like globular clusters or the central bulges of spirals. Ellipticals can be much bigger than spirals. In Figure 14.1 on page 164, NGC4414 has a mass of about $10^{11} M_{\odot}$, while M87 has a mass ten times larger. Moreover, such giant ellipticals often seem to be the places where galactic activity prefers to occur. Quasars (see below) and giant radio galaxies tend to be ellipticals.

For a long time, the contrast between spirals and ellipticals was very puzzling to astronomers. Why should galaxies have formed in two such different ways? How could one explain how a relatively “clean” elliptical galaxy could harbor in its center a massive black hole, and how could it “feed” the hole in order to keep the activity going?

These questions are still open and much debated among astronomers, but the outline of a solution is emerging. The key is galaxy collisions. We see many examples of spiral galaxies colliding with one another, either head-on (Figure 14.3) or in a near miss (Figure 14.4). Computer simulations of what happens when two spiral galaxies collide have been able to reproduce observed galaxies, such as those in the photographs, with remarkable fidelity. The photographs show galaxies that have

collided but have not (yet) merged. However, in a certain fraction of cases, galaxies merge completely. It is now believed that such mergers result in the formation of giant elliptical galaxies.

The tidal gravitational forces associated with the collision compress the gas clouds of the galaxies, triggering star formation on a huge scale, in the way we saw simulated in Figure 12.2 on page 138. Many of these stars evolve rapidly to the supernova stage, and if there are enough supernova explosions in a short time, the remaining gas and dust will simply be blown away. Because the collision usually involves galaxies whose spiral disks were not in the same plane, the final galaxy will not have a disk-like shape any more. It will simply be a highly disturbed collection of stars that eventually relaxes to an elliptical shape.

It may be that in the center of the merged galaxy, the stronger gravity and the initial availability of gas and dust can lead to the formation of a black hole. Or it may be that the original spirals already have black holes in their centers – there is a growing body of evidence that most spirals do, including our own. Again, numerical calculations show that the two black holes can spiral into the center of the new galaxy and even merge together in a relatively short time, perhaps 10^9 years. Issues like this are at the heart of current research into active galaxies.

Galaxies are speeding apart

We can't learn much about galaxies until we establish their distances. Astronomers in the nineteenth century had observed galaxies, but most assumed that they were at the same distance as the ordinary stars of the Milky Way. This made them rather small and insignificant in the grand plan of things. But the American astronomer Edwin P Hubble (1889–1953) changed all that when he determined in 1929 that galaxies were well outside the Milky Way, and were in fact completely separate stellar systems.

Hubble showed more than that, in fact. It was already known from observations of the spectral lines of galaxies that they were receding from us at various speeds. This might, of course, have indicated that they were small systems expelled from the Milky Way by some mechanism, and this is what a large number of astronomers initially believed. But once Hubble had distances, he discovered that the speed of recession of a galaxy was always proportional to its distance from us. We write this *Hubble law* in the following way:

$$v = Hd, \quad (14.6)$$

where v is the recession speed, d is the distance to the galaxy, and H is a constant of proportionality that we now call the **Hubble constant**. Its value, in units that make sense to astronomers, is about $70 \text{ km s}^{-1} \text{ Mpc}^{-1}$. In conventional units, this is about $7 \times 10^{-15} \text{ s}^{-1}$. The accuracy to which it is known is better than 10%.

The Hubble expansion law, plus the large distances involved, demolished the ejection theory and showed that the Universe as a whole is expanding. We shall explore this implication beginning in Chapter 24.

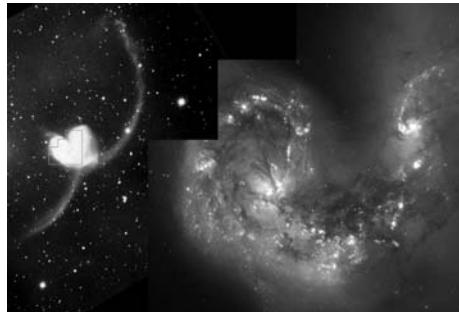


Figure 14.4. A collision between two galaxies that completely changes their appearance. The image is a composite of a ground-based photo (left) and a high-resolution image of the central region taken by the Hubble Space Telescope (right). The galaxies, called NGC4038 and NGC4039, were originally well-separated spirals. Tidal gravitational forces have expelled long streams of gas and stars and stimulated a huge burst of star formation, especially at the join between the two. Photomontage courtesy NASA/STSCI.

In this section: Hubble discovered that distant galaxies are moving away at speeds proportional to their distances from us. This indicates that the entire Universe is expanding.

▷ Astronomers initially called galaxies “nebulae” because they did not know that they were composed of individual stars. Like true nebulae (clouds of gas around stars), galaxies appeared to be just smudges on the sky.

In this section: exploring the Universe depends on knowing the distances to galaxies. Astronomers use a complex hierarchy of distance measures. Only in recent years have they been able to measure the scale of the Universe accurately.

▷ Of course, galaxies also have random velocities on top of the systematic expansion of the Universe. If the galaxy is too near, these will dominate v and make the Hubble method unreliable. In any distance determination, allowance has to be made for this uncertainty.



Figure 14.5. Edwin Hubble's patient measurements on hundreds of galaxies proved first that they were outside the Milky Way, and second that the Universe was expanding and consequently of finite age. Few astronomers have had as a profound an influence on human thought as he. The first Space Telescope was fittingly named after him (HST). Reproduced with permission of AIP Emilio Segrè Archive.

▷ When astronomers refer to "the Galaxy" instead of just "the galaxy" they mean our own galaxy, the Milky Way!

▷ Galaxy and nebula names come from catalog names. The brightest are in the list compiled by the French astronomer Charles Messier (1730–1817); these names begin with M. Many more are listed in the New General Catalog, compiled in Ireland by the Danish astronomer J L E Dryer (1852–1926); their names begin with NGC.

Measuring the Universe: the distances between galaxies

The Hubble Law gives an excellent way of measuring the distance to a galaxy, provided one knows the value of H . The spectrum of light from a galaxy reveals its redshift, from which one deduces its velocity v . Then one just divides v by H to get d . But the central problem of cosmology in the last half of the twentieth century was to determine the value of H . Only since about 1990 have astronomers begun to agree on its value.

Astronomers try to determine H by measuring the distances to some galaxies independently of the Hubble method, and then measuring v to determine H . We shall see below that getting reliable distances to enough galaxies is a very difficult job, and the astronomers' best-guess value of H has changed many times because of this. It is ironic that Hubble's own distances – and hence his own value for H – were systematically wrong. They were the best that could have been done at that time, but the distances came out a factor of five or ten smaller than our present estimates.

The reason for the difficulty is the complexity of the chain of argument that leads to the distances. We described some of the steps in this chain in Chapter 9. Each step requires a standard candle, a class of objects whose intrinsic brightness is known, so that their apparent brightness can be used to measure their distance. An important class of *variable stars* called Cepheid variables can be seen in nearby galaxies, and the orbiting Hubble Space Telescope (HST) has extended observations of them to more distant galaxies. These are useful because their intrinsic luminosity is correlated with their period of variability, so that by timing the regular variations in the star's brightness an astronomer can deduce its luminosity. Cepheids in turn can be used to calibrate other standard candles, such as supernovae of Type Ia, and certain kinds of ionized gaseous regions around hot stars ("HII regions"). As we mentioned in Chapter 12, Type Ia supernovae have become particularly important in recent years because they can be detected so far away that they not only can be used to determine the expansion rate H of the Universe, but also its rate of change, called the acceleration of the Universe. We will see in Chapter 24 and subsequent chapters that these measurements have given the completely unexpected result that the Universe is actually accelerating its expansion.

The size of the Galaxy is much better determined: the Sun orbits the center at a radius of some 8 kpc. The mass of the Galaxy is not so well determined, but it is about $10^{11} M_{\odot}$. Near the Galaxy are several small satellite galaxies, the most prominent of which are the Magellanic Clouds, visible from the Southern Hemisphere. The nearest large galaxy is M31, seen in Figure 9.1 on page 104, which is somewhat larger than the Milky Way. M31 has a prominent satellite elliptical galaxy called M32, whose mass is about $10^{10} M_{\odot}$. M32 is the nearest elliptical galaxy to us, so it has been intensively studied. Despite M32's relatively small size, observations by the Hubble Space Telescope indicate that it probably has a sizeable black hole in its center. M31 and M32 are about 0.5 Mpc from the Milky Way, and falling toward it.

Masses of galaxies and their luminosities are hard to determine. Only in the last two decades have astronomers had instruments, usually in space, that could observe galaxies at most wavelengths, finally getting an estimate of their total emission. The way astronomers measure the masses of the Sun, of planets, and of stars is to monitor objects in orbit about them, and infer their masses from Newton's law of gravity. This will not usually work for galaxies, because orbital times are too long: we can't wait 10^8 years to see what our exact orbital period around the center of the Galaxy is! Of course, if we could measure our acceleration towards the galactic center accurately enough, we could measure the mass of the Galaxy quickly. Only one such measurement has ever been performed, using what astronomers call the

Hulse–Taylor binary pulsar system, whose catalog name is PSR1913+16. We will study this very important system in Chapter 22. For the most part, astronomers use approximate measures of the gravitational accelerations produced by galaxies to measure their masses. We will discuss these when we consider the problem of missing mass in the next section.

When all else fails, astronomers estimate the mass of a galaxy from its brightness. Using a rule of thumb called the **mass-to-light ratio**, astronomers multiply the luminosity of a galaxy by a rather uncertain number to get its mass. The rule of thumb takes account of the fact that much of the mass of the galaxy does not radiate light. Astronomers use values of M/L between 10 and 40 solar masses per solar luminosity. But again, this number is uncertain because of the uncertainties in mass and luminosity of all galaxies.

Most of the Universe is missing!

Given that it is not possible to follow the orbit of a star or satellite galaxy around a galaxy whose mass we wish to measure, and that we cannot determine the instantaneous acceleration of the object, how are we to estimate the galaxy's mass? Gravity is the only reliable way to do it, but we need to make additional assumptions.

Within galaxies, the usual assumption is that stars follow circular orbits. If a star has a speed V in a circular orbit of radius R , then its acceleration towards the center of the circle is (see Investigation 3.1 on page 22) V^2/R . Equating this to the gravitational acceleration of the galaxy (assuming all its mass to be concentrated at its center, at least as a first approximation), GM/R^2 , gives an expression for the mass of the galaxy:

$$M = \frac{V^2 R}{G}. \quad (14.7)$$

Let us try to measure the mass of our Galaxy this way. Astronomers know that the distance R to the galactic center is 8 kpc. But we don't know our orbital speed directly, again because the motion is too slow for us to watch it. Instead, we use a more indirect method. Stars slightly nearer the center should be going faster, because Equation 14.7 says that $V^2 R$ is a constant. By measuring the difference in speed between stars slightly nearer the center than the Sun and those slightly further away, it is possible to determine V itself from Equation 14.7.

Actually, the situation isn't quite so straightforward. For one thing, stars have random motions on top of their orbital motion, so one has to average over a suitable sample of stars in the two different positions. For another, one cannot completely neglect the fact that the mass of the Galaxy is not concentrated at the center: there is mass between the two positions that affects the difference in orbital speeds. The first astronomer to work out how to make a good estimate of the mass of the Galaxy this way was J Oort, whom we met in Chapter 6.

While this method is probably very good, it does rely on the untested assumption of circular motion. If for some reason the orbits of stars near the Sun all follow a single ellipse, on average, then the mass we estimate for the Galaxy will be systematically wrong.

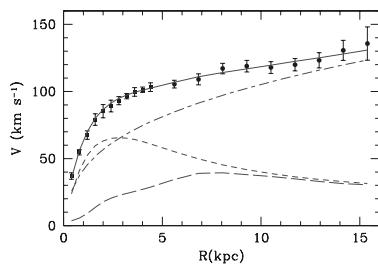
When we apply this circular-orbit assumption to external galaxies, it sometimes is easier to use, and it gives us our first indication of missing mass. If a spiral galaxy is nearly edge-on, then its circular orbits will be moving directly towards us or away from us in certain places. These velocities produce Doppler shifts in spectral lines, and so by measuring these shifts one can deduce the speed of matter in the galaxy in different places. If we average the speeds measured at, say, 5 kpc to one side and to the other of the center, we should obtain the overall velocity of the galaxy away from (or towards) us. If we take the difference of these speeds, we should obtain

In this section: the central problem of understanding galaxies today is that there appears to be much more dark matter than luminous. This is not concentrated in regions where galaxies emit light, but is spread outside and between them.

$2V$. Knowing the distance to and size of the galaxy (not always easy!) gives us the orbital radius R , and hence the mass M .

If the gas or stars whose motion we measure in this way are far enough out from the center of the galaxy, then one would expect that the mass inside the orbit would be fairly constant. Then, by Equation 14.7, one would expect to see V^2R constant, or $V \propto R^{-1/2}$. What we actually see in almost every case where measurements can be made far from the galactic center is illustrated in Figure 14.6: V stays relatively constant or even increases as R increases.

Figure 14.6. The orbital speed of gas in the spiral galaxy M33. If the mass of the galaxy were all in its center, then we would expect V to decrease. Instead, it is slowly increasing as far out as it can be measured. The two lower dashed lines are contributions to the velocity from the visible galaxy. The upper dashed line is the deficit that must be made up by invisible mass. Figure from E. Corbelli and P. Salucci, Mon. Not. Roy. astr. Soc., 311, 441 (2000), with permission of the authors.



the photographic image.

What is this missing, or dark, matter? No one yet knows, despite years of investigation. We will return to this question after the next section, once we have seen that the spaces between galaxies hide even more missing mass than the galaxies themselves do.

Gangs of galaxies

In this section: galaxies often come in groups called clusters, with hundreds or thousands of members.

These clusters provide additional evidence for missing matter, and they also give clues to how galaxies were formed in the very early Universe.

Figure 14.7. The central portion of the Virgo Cluster of galaxies. The elliptical galaxy M87 illustrated in Figure 14.1 on page 164 is at the center of this picture. Several hundred galaxies of various sizes may belong to this moderately rich cluster. Use of this image is courtesy of the Palomar Observatory and Digitized Sky Survey created by the Space Telescope Science Institute, operated by AURA, Inc. for NASA and is reproduced here with permission from AURA/STScI.

The ever-attractive nature of gravity makes it inevitable that galaxies are not spread out uniformly through the Universe. Instead they tend to group together in what are called clusters of galaxies. Our own Local Group is a small, loose cluster consisting of 3 spiral galaxies (Andromeda, the Milky Way, and a smaller spiral called M33, illustrated in Figure 14.6), several minor galaxies, and many satellite galaxies such as M32 and the Magellanic Clouds.



What are we to make of the **rotation curve** shown in this figure? If V is constant, then Equation 14.7 on the previous page tells us that the mass of the galaxy inside a distance R is proportional to R . Now, the rotation curves obtained in this way use very weak radio waves from neutral hydrogen gas orbiting the galaxy. This is sometimes detectable two or three times as far from the center as any visible light from the galaxy, in other words in regions well outside the photographic image of the galaxy. They tell us that the mass is still increasing: there is a huge amount of dark matter out there, perhaps two or three times as much mass as one would infer from

There are many clusters that are much more populous, having 100 to 1000 members. The nearest big cluster of galaxies is the Virgo Cluster, at a distance of about 18 Mpc, containing a few hundred galaxies. Figure 14.7 is a photograph of this cluster. The elliptical galaxy M87 shown in Figure 14.1 on page 164 is a giant elliptical in the center of the Virgo Cluster: it may well have been formed by the merger of two or more spirals that were brought to the center by the collective gravitational force of the whole cluster. Clusters also group into **superclusters**; we will discuss these in Chapter 25.

When we try to estimate the masses of clusters, we find further evidence of even

more missing mass. The mass of a cluster has to be inferred from its gravity, and there are several ways to do this. We shall consider three of them.

One way is to estimate the mass from the observed velocities of cluster galaxies. This is called the **virial method**, and it is a generalization of simple ideas we have seen before. A planet in a circular orbit around a central star has a kinetic energy that is exactly one-half of the energy it needs to escape from the star. The same factor of one-half applies in more complex systems involving many orbiting bodies, provided that the cluster of galaxies is relaxed, in the sense we defined for globular clusters above. In a cluster of galaxies, no single galaxy has a nice circular orbit about the center, but the *average* energy of all the galaxies at some position will be the same as the energy of a circular orbit there. And the total kinetic energy of all the galaxies will be just half of the “escape energy”, which for a cluster is the energy required to break it up, its binding energy. This is the energy we looked at in Investigation 14.1 on page 165.

To use this, astronomers try to measure the kinetic energies of all the galaxies and then equate the total to half the binding energy. The binding energy depends somewhat on the distribution of galaxies (how concentrated towards the center they are), but it is proportional to GM^2/R , as in Investigation 14.1 on page 165. In principle, by measuring the distribution and velocities of galaxies and determining the distance to the cluster and hence its radius R , one can infer M . There are many uncertainties in doing this, not the least of which is the fact that it is hard to be sure that any given galaxy is a member of a cluster. On top of this, many clusters are not really relaxed. Nevertheless, the method indicates that the true mass of a cluster can be up to 20 times the visible mass of the galaxies. Interestingly, the first application of this method, and therefore the first demonstration of the missing mass, was by the Swiss physicist and astronomer Fritz Zwicky (1898–1974) in 1933. Since even the understanding that galaxies were external systems was only four years old at that time, astronomers did not know what to make of this result, and simply ignored it for decades! We will learn more about Zwicky’s far-sighted research in Chapter 20 and Chapter 23.

Our second way of estimating cluster mass is another form of the virial argument, made possible by X-ray observations. Somewhat to the surprise of most astronomers, X-ray telescopes have detected strong X-ray emission from the spaces *between* galaxies in most dense clusters. This indicates that there is a hot gas in this space. This gas seems, from the observations, to be fairly smoothly distributed, and so it is almost certainly relaxed. The gas allows astronomers to estimate the mass of the cluster. Doing it crudely, we just equate half the mean kinetic energy $3kT/2$ of an atom of the gas (which is mainly hydrogen) to its gravitational potential energy, $GM_{\text{cl}}m_p/R_{\text{cl}}$. A rich cluster might contain 500 galaxies in a region of radius 3 Mpc, and its gas might emit X-rays with an energy that reveals that the gas temperature is 50 million degrees kelvin. Simple arithmetic allows us to solve for M , which comes out near to $10^{15} M_\odot$, about 20 times the visible mass of the 500 galaxies. This is consistent with the numbers found by the virial method on galaxy speeds. Note that this is an estimate of the total mass of the cluster, not just the mass of the gas generating the X-rays.

The third method, which is unrelated to virial or dynamical estimates, is to use gravitational lensing, which we will study in Chapter 23. According to general relativity, the direction light moves is deflected when it passes any gravitating body, and the consequences of this are often easily seen in astronomical photographs. When light from a distant quasar or galaxy passes by a galaxy cluster on its way to us, then astronomers can estimate the mass that the cluster must have in order to produce

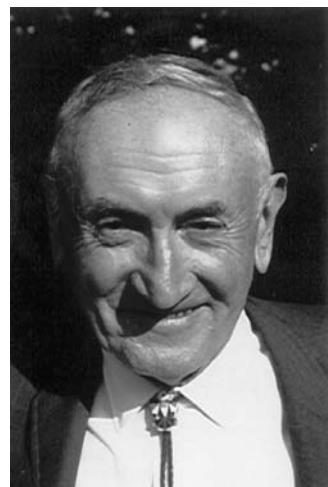


Figure 14.8. Fritz Zwicky was one of the most creative, but also one of the most idiosyncratic, astronomers of the mid-twentieth century. Trained as a theoretical physicist, he moved to the California Institute of Technology in 1925 and remained there until his retirement in 1968. Observing with some of the world’s most powerful telescopes, he used his physics background to make deeply perceptive interpretations of his data. He was the first to recognize the problem of the missing mass, the first to suggest that supernovae produced neutron stars, the first to face the likelihood that black holes could form, and the first to suggest that gravitational lensing would be observable. All of these ideas became mainstream astronomy, but in most cases only after his death. Perhaps because his contributions were so often ignored or undervalued by his contemporaries, Zwicky developed a reputation of being short-tempered, impatient, and critical toward other astronomers, although to students he was frequently welcoming and encouraging. He also developed a theory of mental processes called morphology, which has a small following today. The photograph shows Zwicky in 1970 (reproduced courtesy of the Fritz-Zwicky-Stiftung, Glarus, Switzerland).

the observed deflection. This usually leads to estimates a little larger than those of the virial and X-ray methods. Interestingly, this method is beginning to be used to detect dark clusters, regions that produce gravitational lensing but contain no visible galaxies or X-ray emission at all!

Some of the mass of clusters that is not in the galaxies is the X-ray emitting gas itself. This gas is presumably left over from the time the galaxies formed, so it might be thought to be “primordial”, i.e. not processed through stars. But studies of the spectra of X-rays from the gas indicate that it contains many elements that can only have been formed in stars, such as iron. This indicates that the dynamics of clusters is complicated, and that gas moves into and out of galaxies over time.

The intensity of the X-ray emission allows one to estimate how much gas there is. Compared to the mass in the galaxies, the cluster gas dominates: there is perhaps five times as much mass in the diffuse gas as in the galaxies. But this is not enough to account for the total mass of the clusters: the X-ray gas amounts to only about 25% of the total mass of a typical cluster. The rest is dark, emitting no visible light, no X-rays, no infrared light, no radio waves.

The missing mass

In this section: something must be between galaxies that is providing the background gravity but is not participating in the gas dynamics that leads to the emission of light,

X-rays, radio waves, or infrared radiation. This is the dark matter.

Its identity is not known. The puzzle of dark matter is one of the most important in all of astronomy and physics.

The missing cluster mass may be the same stuff as is missing in galaxies themselves; there is more of it because the spaces between galaxies are so much larger. What form the dark matter takes is as uncertain for clusters as it is for individual galaxies. Possible explanations fall into two groups.

First, it could consist of objects of astronomical size, such as very small stars or large black holes. The mini-stars are called **brown dwarfs**. They are not massive enough to ignite nuclear reactions, so they just quietly contract, shining weakly by radiating their gravitational potential energy away. We know that such a star must be dim, since we showed in Investigation 11.1 on page 123 that gravitational potential energy would not sustain the Sun’s luminosity for more than about 0.1% of its lifetime. Therefore, the average luminosity of a brown dwarf over several billion years must be less than $0.001L_{\odot}$. Such stars are not easy to detect.

Black holes in the dark matter might be left over from a hypothetical early burst of star formation and collapse that occurred as the galaxy formed. Many scientists believe that the first generation of stars, made of pure hydrogen and helium, would contain more massive stars than formed later, and many of these might have formed black holes.

These two populations would be difficult to observe directly, but might show up in studies of gravitational lensing within the Galaxy. Teams of astronomers are systematically monitoring millions of stars in the hope of finding chance moments when one of these hypothetical dark objects passes close enough to the line-of-sight to one of the stars to magnify its intensity briefly. The good news is that these studies have uncovered a large population of dark stars with a mass apparently between 0.1 and $0.5M_{\odot}$, which are called MACHOS. But the bad news is that they are not plentiful enough to account for most of the missing mass of our own Galaxy, let alone that of most clusters.

A second possibility, and the one that is favored by most physicists, is that the missing mass could consist of a smooth distribution of some kind of elementary particle. If so, the particles cannot carry electric charge, for otherwise in the X-ray emitting gas in clusters the “dark” particles would collide with the ordinary gas and emit X-rays themselves, creating more intense emission than is observed. So the particles must be electrically neutral. The problem is that physicists don’t know of any neutral particles that have enough mass to provide the extra gravity and are stable against radioactive decay. A free neutron, for example, decays into a

▷ MACHO stands for MASSIVE compact halo object.

proton, electron, and neutrino in only 11 minutes. Dark matter has to last for more than 10 billion years! Neutrinos have mass and, collectively, are stable, but their masses are too small. Not only would the dark matter require more neutrinos than physicists believe could ever have been produced, even in the early Universe, but considerations of how galaxies formed in the first place require particles that have at least hundreds of times the neutrino mass.

Studies of galaxy formation lend considerable support to the dark matter hypothesis, despite the fact that the dark particles are not known. We will come back to this issue below, and in much more detail in Chapter 25, where we can study the formation of galaxies in the context of the overall expansion of the Universe. For now, we will just note that most physicists believe that galaxies did not form spontaneously, but rather that they needed “seeds”: strong centers of gravity to start pulling the gas in until it reached a high enough density to make the gas collapse inwards and begin forming large numbers of stars.

The seeds could have come from these dark matter particles, provided the particles had high enough mass and weak enough interaction with ordinary matter (with the protons and electrons) to cool off rapidly as the Universe expanded, to clump, and to pull in the ordinary gas. This hypothesis, called “cold dark matter”, has been studied in numerical simulations on large supercomputers, and it seems to produce galaxies with the size, number, and distribution that astronomers observe. So the cold dark matter (CDM) hypothesis neatly explains both the missing mass and the formation of galaxies.

The elementary particles in this picture are sometimes called WIMPs. While they are not predicted by the standard theories of particle physics, they could plausibly emerge from unified theories of the nuclear, weak, and electromagnetic forces. All that is required now to turn CDM from theory into fact is to observe WIMPs directly. Millions of them must pass through us every second. Like neutrinos, they are very elusive and almost never collide with atoms of our bodies. Experimental searches for WIMPs are now underway; if enough exist to account for the missing mass, they ought to be detected and identified in the near future.

There are other possible seeds for galaxy formation. Among the best-studied alternatives are **cosmic strings**. These are long, incredibly thin concentrations of mass that grow in the early Universe in some theories of high-energy physics. They could also provide sites where galaxies form. Simulations of galaxy formation do not show such good agreement with the distribution of galaxies that we see today, and cosmic strings would not easily provide the missing mass within galaxies, since they are typically much longer than the spaces between galaxies. For these reasons, most astronomers today favor WIMPs over cosmic strings.

Although it is rather embarrassing for astrophysicists to have to admit that they do not know what most of the mass in the Universe is, the missing mass problem is one of the crown jewels of modern astronomy. Much of astrophysics deals with the application of known laws of physics to try to model and understand astronomical phenomena. But occasionally astronomy offers the only way for a new discovery to be made in fundamental physics. Newton used astronomy to determine his law of gravitation. Studies of solar neutrinos have revealed the striking phenomenon of neutrino oscillations. Gravity, by telling us what the masses of galaxies really are, appears now to be pointing the way to further new physics. The new particles do not fit into the known theory of the nuclear forces. They would be indicators of physics outside of this theory, and most physicists would expect them to be vital evidence for the theory that will unify all of the forces of physics. We will return to this issue in Chapter 27.

▷WIMP stands for weakly interacting massive particle.

▷The evidence at present favors the WIMPs over the MACHOs!

If there is new physics to be discovered here, it would not have been found except for the painstaking study of the dynamics of galaxies. We will see a further modern example of how astronomy can be used to discover new physical laws when we discuss the way elements were formed in the Big Bang (Chapter 25).

Radio galaxies: the monster is a giant black hole

In this section: radio galaxies emit radio waves from regions far outside the visible galaxy, powered by jets of gas emitted from their central regions. The sources seem to be giant black holes formed from millions or even billions of solar masses.

Figure 14.9. Radio emission from M87 is aligned with the jet in the inner region and then spreads out and changes direction. Since the galaxy is more than 10 kpc in radius, or roughly 30 light-years, the large region of emission outside the galaxy indicates that the jet has been active for hundreds of thousands of years. Radio image by Owen, Eilek, and Kassim at the Very Large Array in New Mexico.



Radio emission from galaxies is generally associated with jets of gas streaming outwards from a central black hole at nearly the speed of light. Figure 14.9 shows this for the elliptical galaxy M87, which we have seen in previous illustrations. The radio emission comes from a large region surrounding the galaxy, which is coincident with the brightest part of the radio image. In the inner regions it is aligned with the jet we saw in Figure 14.1 on page 164. Notice that the radio emission goes in both directions from the center, which means that there is probably a jet in both directions, even though only one is visible in Figure 14.1. As the jet leaves the galaxy, the radio emission pattern makes a turn. This could indicate that the jet is running into gas outside the galaxy and is being deflected. The size of the radio lobes indicates that the activity has been taking place for at least hundreds of thousands of years.

These features are absolutely typical of giant radio galaxies. Indeed, M87 is a baby among them: the most luminous ones are thousands of times as bright and ten times the size. Their activity has been going on for millions of years.

What are we to conclude from this? The only mechanism available to a galaxy for maintaining a single direction steady over such a long time is rotation: a rotating disk of gas and/or stars will define an axis of rotation that can normally remain fixed for very long times. Moreover, the dynamical studies of the inner region of M87 indicate that there is a black hole there. Presumably this is not a coincidence.

How does the black hole generate the jets? Where does the energy come from, for example? Nuclear energy is simply not adequate. Consider the numbers: many radio galaxies radiate 10^{38} J s^{-1} in radio waves, which is ten times as much as a typical galaxy radiates in optical light. Yet, as we see in the pictures, the jet originates in a tiny region in the center. No set of nuclear reactions such as we described in Chapter 11 for normal stars could produce this prodigious energy. One has to think of mechanisms for converting the mass of whole stars into energy. The radio luminosity above is the equivalent of converting $1/60^{\text{th}}$ of the mass of the Sun into pure energy every year, using Einstein's famous equation $E = mc^2$, which we will study in the next chapters. And this conversion process must be sustained for millions of

years.

Gravity in relativistic situations offers ways of doing this. A particle falling onto a neutron star reaches it with a speed equal to the escape velocity. We saw in Chapter 12 that the escape velocity from a neutron star is about half the speed of light. Its kinetic energy at this velocity, $mv^2/2$, is therefore a good fraction of its total rest mass energy mc^2 . This energy is in principle available to any processes near the neutron star that could convert it into power for a jet.

In fact, neutron stars are much too small to act as centers for the jet phenomenon. If, say, 20% of the infalling mass is converted into jet energy, the remaining 80% of the mass has to stay near the neutron star, since to send it back out would take as much energy as the mass released by its falling in. If the jet requires the conversion of 1/60th of a solar mass in energy each year, then 1/15th of a solar mass per year must accumulate on the star. After only something like 15 years, this would push the neutron star over the upper mass limit and convert it into a black hole. Over a few million years, at least $10^5 M_\odot$ will have accumulated in the region where the jet originates. Therefore, the mechanism needs a massive central black hole for its relativistic gravitational field. Astronomers call this massive black hole "the monster".

Other possibilities have been proposed: supermassive relativistic stars, extremely dense clusters of neutron stars, and others. It seems, however, that even if one could somehow form such systems, they would not last long before collapsing to form a massive black hole. The conclusion that the monster is a massive black hole seems inescapable.

Quasars: feeding the monster

The discovery of the enormous luminosity of quasars in 1963 by the Dutch astronomer Maarten Schmidt (b. 1929) was a landmark in the development of modern astronomy. Radio astronomers had identified a class of unusual, intense radio sources that did not seem to be associated with galaxies. Optical observations revealed point-like images at the positions of some of the radio sources, but the images were not like ordinary stars. In particular, their spectra did not look like spectra of stars, and in fact no-one could identify any of the lines. They were called *quasi-stellar objects*, a name that has evolved into *quasar*, and is frequently abbreviated QSO.

Schmidt decided to see if he could interpret the spectrum of one of these objects, called 3C273, by applying a very large Doppler shift to some standard spectral lines of hydrogen. He found that he could indeed fit the spectrum of 3C273, provided he used a shift corresponding to a recessional velocity of 15% of the speed of light. This was far larger than any velocity that had by then been measured for galaxies. Interpreted as a Hubble velocity, it meant that 3C273 was one of the most distant objects known. Although it looked like a dim star on photographic plates, its great distance meant that it was actually one of the most luminous objects known.

The redshifts of other quasars were soon measured, and a number of features emerged. First, quasars put out much more light than an ordinary galaxy, so it is likely that they are associated with some phenomenon that takes place in the center of some galaxies, like that which produces radio galaxies. Second, most quasars are so far away that the light we see has been traveling to us from them for a good fraction of the time since the Big Bang. The Universe was younger then, and it is natural to conclude that quasars have something to do with the early stages of the formation of their "host" galaxies. Third, quasars were much more numerous in the early Universe than they are now. They were so numerous that a sizeable fraction of galaxies must have had quasars in them at one time, although in our Galaxy's

In this section: quasars also seem to contain black holes, and they give a clue to the source of the energy: gas falling towards the black hole.

>Initially there was some skepticism that the enormous quasar velocities should be interpreted as part of the Hubble expansion. But very sensitive optical observations have revealed many quasars in clusters of normal-looking galaxies of the same redshift. There can therefore be no doubt about the enormous distances to these objects, and hence their enormous luminosities.



Figure 14.10. The jet from QSO 3C273. It comes straight out of the center of the object. The jet illustrates the close similarity between quasars and radio galaxies. Photo courtesy

AURA/NOAO/NSF.

neighborhood they seem to have died out completely.

The key observation that shows that quasars are related to radio galaxies is that they also display jets. In Figure 14.10 we see the jet from the original quasar, 3C273. It comes straight from the heart of the image.

Quasars give us, in addition, some very important information on how big the region emitting their radiation can be. The brightness of quasars is very variable, and some have been known to change their brightness by a factor of two in a few minutes! Whatever the mechanism for changing the luminosity may be, relativity tells us that it cannot involve influences that move faster than light, so the size of the emitting region must be smaller than the distance light can travel during the time that the luminosity changes. In this case, this is a few light-minutes, or less than 1 AU, less than the distance of the Earth from the Sun.

All the light of a quasar, more than the normal emission from a whole galaxy, originates in a region smaller than our Solar System!

This is consistent with the conclusions we came to for radio galaxies. Even a huge black hole of $10^9 M_\odot$ has a size of $2GM/c^2 = 20$ AU (recall the formula for the size of a black hole, Equation 4.12 on page 36), so there is plenty of room to fit such a hole in the monster's chamber! However, there is not much room for any other kind of object that could produce the light from a quasar.

Once quasars were discovered, it became possible to look for them using optical images, even without radio positions. It has emerged that only about 20% of quasars actually emit detectable radio energy. Moreover, similar objects, on a somewhat smaller scale, have been found in the centers of ordinary galaxies. These *active galactic nuclei* come in a wide variety of forms, and classes of them have special names: Seyfert nuclei and BL LAC objects are two. Because quasars are so bright, it is difficult to see their surrounding galaxy, but a few relatively nearby ones have been detected (including that of 3C273). They are all giant ellipticals. The active galactic nuclei, however, can be present in both spirals and ellipticals.

In fact, it now seems that *most* galactic nuclei show some modest level of activity. This means that by observing nearby galaxies, we have a chance of seeing details of the phenomenon that we would never be able to resolve in distant quasars. Observations of M87 using the Hubble Space Telescope have shown that its jet originates in a region that contains a disk of orbiting gas that is no larger than 20 pc and whose orbital speed is at least 750 km s^{-1} . This speed is 25 times the orbital speed of the Earth around the Sun, yet the orbiting gas is 4 million times further from the center of its orbit than the Earth is. Since the orbital speed of a planet is $v = (GM/R)^{1/2}$, the central mass M is proportional to Rv^2 . It follows that the central mass is 2.5×10^9 times as large as the Earth's central mass, which is of course the Sun. For an object of this mass to be smaller than 20 pc in size, it must be a black hole: no method of concentrating this much mass in that small a region could avoid gravitational collapse for long. This is the biggest monster for which astronomers have such conclusive evidence at present.

But we still have the question: what causes the phenomenon? Although the question cannot yet be answered in full, quasars and active galactic nuclei bring the answer closer. Why should the quasar phenomenon have died away with time? How can M87 have such a huge black hole and yet be only modestly active, compared to quasars? And how can active black holes be responsible for the production of jets: don't they trap everything that falls into them?

The final question holds the key. Black holes provide the gravitational attraction that allows matter falling towards them to release such huge amounts of energy, but

it must be that the material is stopped, or at least slowed down, before it actually reaches the hole. If infalling material has some rotation, then it will form an accretion disk around the hole, and it will only gradually spiral in. Before it reaches the hole it will have released a lot of energy. When it falls into the hole, it makes the hole rotate.

How the jet is produced is very unclear at present. It might be that the accretion disk is very thick near the hole and only allows a small opening along the rotation axis, through which matter is expelled by the complicated pressure forces in the disk. Alternatively it is also possible that magnetic fields generated by the matter in the disk could interact with a rotating black hole to generate a jet.

Given that the monster has to be fed by gas falling into the accretion disk, the decay in quasar activity with time could be explained by famine: in the original galaxy there are only a certain number of stars that are in orbits that take them close enough to the black hole to be disrupted by its tidal forces and end up in the disk. Once these have been eaten, the hole becomes quiet. When galaxies merge, stellar orbits become disturbed and the quiet holes suddenly have much more to eat again, and activity can start up again.

Galaxy formation: how did it all start? Did it all start?

Quasars are associated with galaxies that are young. Astronomers have been conducting intensive searches to find galaxies at an even earlier stage, when they are forming, presumably by the contraction of a diffuse cloud of gas in the very early Universe. They have found very interesting objects, very distant, very young. Figure 14.11 shows a sample of these objects found by the Hubble Space Telescope. They are unlike any galaxies we see today. They are mere fragments. It seems clear from studies like these, as well as from numerical simulations using supercomputers, that galaxies form from multiple mergings of such fragments.

Observations of the **cosmic microwave background radiation**, which we will study in Chapter 24, are also beginning to yield information about the mechanisms of galaxy formation (Chapter 25). The initial density irregularities that grew to form galaxies have left imprints on the background radiation, and measurements during the first decade of the twenty-first century should reveal much of the detail of the earliest phase of structure formation in the Universe.

To understand this phase, we need to study cosmology, the history of the Universe in the large. Galaxy formation is but one step in a long chain of events that led to the Universe we observe: normal matter (protons, electrons, neutrons) first took form, hydrogen and helium were made from these building blocks, the dark matter began to clump, hydrogen and helium were drawn in and began forming stars, galaxies, and clusters. After that, we have already drawn the outline of the remaining steps: stars made the heavier elements, the Sun formed from a cloud of gas with plenty of heavy elements, the Earth formed from these heavy elements, and (leaving out a few more steps) here we are!

But wait – before going down this road, how can we be sure that there was a time when galaxies were young, when stars were just beginning to form? We saw in Chapter 11 that the Universe cannot be infinitely old, because stars systematically use up the hydrogen and make more and more heavy elements. There is an even more dramatic way of seeing that the Universe had to have a beginning, a way that

In this section astronomers are beginning to probe the time when galaxies formed, and they see them arising from mergers of smaller objects. We have known for a long time that there was a time when galaxies formed. Otherwise the sky would not be dark at night. This is Olbers' Paradox.

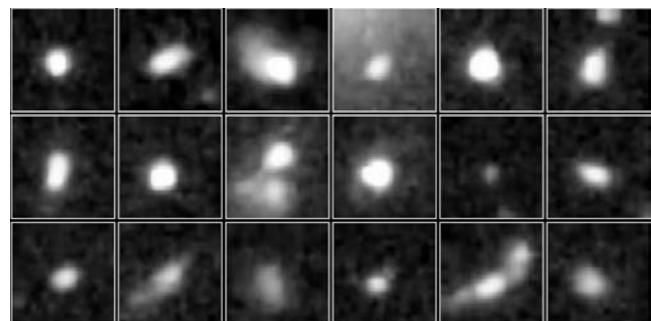


Figure 14.11. A Hubble Space Telescope collection of images of very distant, very young clumps of stars that are in the process of combining to form galaxies as we know them. Taken from STSCI PRC96-29B, images by R Windhorst, NASA/STSCI.

was already discussed in Newton's day. It is what we now call **Olbers' Paradox**.

Look up at the night sky on a clear night. It is dark. That is enough to show that the Universe had a beginning!

The argument is as follows. Let us assume a model of the Universe that would have seemed very reasonable to eighteenth century physicists: the Universe is filled with a uniformly distributed collection of stars, and is infinite in extent, and infinitely old. This was, in fact, the model favored by Newton. Then the problem is that the light radiated by stars would also be uniformly distributed in such a Universe. Moreover, since energy is conserved, the light radiated by stars long long ago is still running around the Universe, and since this has been going on for an infinite amount of time, the amount of light at any location in the Universe would be infinite. The night sky would not only be bright, it would fry us!

▷ Put mathematically, the light from any star at a large distance r from us is dimmer than that from a star at the nearer distance R by the ratio $(R/r)^2$, but a spherical shell surrounding us with thickness δr has a larger volume if it is at a distance r than if at R , by the inverse factor $(r/R)^2$. Thus, the number of stars (in such a uniform model) is exactly large enough at r to provide the same amount of light here as the stars at R . Since there are an infinite number of such shells, there is an infinite amount of light.

This argument was put to Newton by the English physician William Stukeley (1687–1765), but Newton ignored it. The German astronomer Wilhelm Olbers (1758–1840) took the issue seriously, and tried to find a way out. He argued that the light might be absorbed by dust or by other stars. But this does not work: it just leads over time to the heating of the absorber, which will then re-radiate the light. Another way out would be to assume that the part of the Universe containing stars is of finite extent, surrounded by empty space, so that the light reaching us is finite as well; the light radiated by stars long ago has left our part of the Universe and is streaming out through the empty space around it. But this is no solution either: a finite Universe must collapse in on itself through Newtonian gravity, and so it would have only a finite lifetime, as well as a finite history. Newton himself apparently believed that God would intervene periodically to stop the Universe collapsing on itself, but he did not postulate a Universe of finite extent.

That no-one before Einstein seriously discussed a Universe of finite age as a resolution of Olbers' Paradox illustrates the fact that, until Einstein, few scientists seriously thought that cosmology was a province for purely scientific investigation. All that has changed, as we will see in Chapter 24. Today, cosmology is the deepest application of gravity to science. But to understand it we need to understand Einstein's gravity, general relativity. So we shall defer our final investigation of galaxies, how they formed, and how they are distributed in the Universe, until after we have learned how Einstein re-formulated Newton's concept of gravity.

Physics at speed: Einstein stands on Galileo's shoulders

We have allowed gravity to take us on a tour of the Universe in the first half of this book. It has taken us from the planet Earth to the rest of the Solar System, then to other stars, and from there to galaxies. Gravity wants to lead us further, because we have not yet come to understand its most profound consequences. These include black holes, which we met briefly in Chapter 4, and the Big Bang, which is the beginning of time itself.

These matters require *strong gravity*: gravity that is strong enough to trap light in a black hole or to arrest the expansion of the entire Universe. Studying strong gravity takes us beyond the limits where we can trust Newton's theory of gravity and his laws of motion.

The reason is speed: if gravity is strong, then speeds get large. If we shrink a star until it is compact enough to turn it into a black hole, then the escape speed from its surface gravitational pull will exceed the speed of light; all other speeds near it, such as the speed of a nearby orbiting planet, will be close to the speed of light. But when we try to understand phenomena that involve speeds close to that of light, we need relativity. We need to improve our theory of gravity to make it a *relativistic* theory.

Einstein showed physicists and astronomers how to do that. He did it in two steps. First he showed how physics gets modified when things move at close to the speed of light; this came to be called the special theory of relativity, or special relativity. In this work, he basically brought Galileo's principle of relativity (recall Chapter 1) up to date. The second step came ten years later, when he brought Newton's gravity up to date with relativity. Einstein's relativistic theory of gravity came to be called the general theory of relativity, or general relativity.

"If I have seen farther," Newton once said of his relationship to scientists of an earlier age, "it is by standing on the shoulders of Giants." The same could be said of Einstein: his special relativity rests squarely on Galileo's shoulders, and he made his general theory of relativity as close to Newton's theory as he could.

Fast motion means relativity

Unfortunately, special and general relativity are probably the worst-named theories of modern physics. Their names convey little meaning, and this sometimes causes confusion right from the start. Here are thumbnail definitions of what the theories are really about.

Special relativity is Einstein's description of how some of the basic measurable quantities of physics – time, distance, mass, energy – depend on the speed of the measuring apparatus relative to the object being studied. It shows how they must change in order to guarantee that Galileo's principle of relativity (that the laws of physics should be the same for

In this chapter: we embark on relativity. We present the fundamental ideas of special relativity. Einstein based it partly on Galileo's relativity, partly on a new principle about the speed of light. We discover the main consequences of the theory, which we require for the development of general relativity in the rest of this book.

►The picture behind the text on this page represents the products of the head-on collision of two protons in the CDF experiment at Fermilab, a major particle-physics accelerator laboratory in Illinois. Hundreds of sub-atomic particles are produced in each collision. Such collisions illustrate many of the predictions of special relativity detailed in this chapter. The protons collide at nearly the speed of light, with mass-energy nearly 1000 times larger than their rest-masses. This mass-energy is converted into the rest-masses and energies of the product particles. Accelerator experiments like this test special relativity stringently millions of times per day. This image is captured from the live display of results on the experiment website. See <http://www.fnal.gov>.

In this section: relativity theory is required when motions can be a significant fraction of the speed of light. Gravitational fields that are strong enough to accelerate bodies to such speeds must be described by a theory of gravity that is compatible with the principles of relativity.

every experimenter, regardless of speed – recall the discussion in Chapter 1) should hold even at speeds near that of light.

Because it deals with the general properties of measurements, special relativity is not really a theory about any particular physical system. Rather, it is a set of general principles that all the other theories of physics have to obey to deal correctly with fast-moving bodies. All the theories of physical phenomena – for example, mechanics (the theory of forces and motion), electromagnetism (the theory of electricity and magnetism), and quantum theory (the theory of the sub-microscopic physics of electrons, protons, and other particles) – have relativistic versions that physicists use when they need to understand a situation where speeds get close to that of light.

Gravity is no exception: it must also follow the principles of special relativity.

General relativity is Einstein's relativistic theory of gravity.

General relativity replaces Newton's theory, which works well for slow motion and is still good for most purposes when describing Solar System orbits and the structures of stars and galaxies. But when gravity is strong enough to accelerate bodies to nearly the speed of light, then we have to turn to general relativity to find out what really happens. In Chapter 19 we will see by explicit calculation that Newtonian gravity cannot predict correctly the gravitational field of a moving body.

In what situations do we require a relativistic theory of gravity? One answer – Einstein's answer – is that we need it everywhere, because it is simply unacceptable to have theories of physics that are inconsistent with one another. In Einstein's day this was the *only* answer, since there were no big observational problems with Newtonian gravity. But today we can see much more of the Universe, and there are places where Newtonian gravity simply will not work satisfactorily. So today there is the astronomer's answer, which is that we need to do better than Newton whenever gravity is strong enough so that speeds within the system being observed are near the speed of light. These speeds could be escape speeds or speeds of random motion of gas particles; in the latter case the pressure becomes large, so that p is of the same order of magnitude as ρc^2 .

We saw in Chapter 4 and Chapter 6 that, in Newtonian gravity, the escape speed from a body of mass M and size R is $(2GM/R)^{1/2}$. We can make this large by either increasing M or reducing R , or both. Black holes and neutron stars normally form by reducing R : the nonrelativistic inner core of a star collapses, raising the escape speed until we need relativistic gravity to describe it. Cosmology, the study of the Universe as a whole, needs a relativistic description because it involves a large mass M . If we imagine a region of the Universe with a uniform average mass density ρ , then the mass in a sphere of radius R is $M = 4\pi\rho R^3/3$, so the escape speed $(2GM/R)^{1/2}$ from that region is

$$\text{escape speed from region of size } R \text{ and density } \rho = (8\pi G \rho R^2/3)^{1/2}. \quad (15.1)$$

►This ρ is obtained by spreading the mass of stars and galaxies smoothly over the entire region; it is therefore much less than the density within a star.

This increases in proportion to R , so if we take a big enough region of the Universe we are bound to reach a point where the Newtonian escape speed would be c , and therefore the Newtonian description of gravity would fail. It is precisely on this length-scale that we need general relativity to provide us with a consistent model of the Universe.

Before we can run, we must walk; before we can understand the Universe, we must learn relativity. The present chapter opens the door to relativity. It introduces the basics of special relativity, covering all the essential ideas and the few formulas that we will need when we go on to general relativity, black holes, and cosmology.

Special relativity is both fascinating and – let us admit it right away – worrying. It insists that we change the notions of time and length that we have taken for granted all our lives. This insistence fascinates some people who meet the theory for the first time, and it raises resistance in others: “How can that *be?*” is the frequent question. For readers who want to go deeper into the theory, or who really need to find out how it can *be!*, I have developed the themes covered in the last part of the present chapter more fully in the next, Chapter 16. Reading the next chapter is not essential: if all you want to do is to get on to black holes, then you can jump to Chapter 17 at the end of this chapter.

Relativity is special

Although the name “special relativity” is not particularly informative, it does remind us that Einstein built it on the foundations of Galileo’s principle of relativity, which we met in Chapter 1. Relativity is a thread that has run through physical thinking for centuries, and Einstein reminded his generation of physicists how important it was. In his paper on special relativity in 1905, he built a revolution in physics on the very traditional foundation of the principle of relativity.

Recall the principle of relativity as we phrased it in Chapter 1:

Relativity: All the laws of physics are just the same to an experimenter who moves with a uniform motion in a straight line as they are to one who remains at rest.

Einstein insisted that this was one of *only two* guiding principles that all theories of physics had to follow. He was firmly in Galileo’s tradition here, but it happened that nineteenth-century physicists had by and large begun to doubt that this principle was correct. Einstein rescued the principle of relativity from oblivion.

Einstein’s second guiding principle was far from traditional. In fact, it was so radical that even today beginners have difficulty believing that it can be true.

Speed of light: It is a fundamental law of physics that the speed of light has a particular, fixed value, which we usually call c . Because this is a law of physics, it follows from the principle of relativity that every experimenter who measures the speed of light will get the *same* value for it. This is true *even if two different experimenters moving relative to each other measure the speed of the same beam of light*.

We are used to the speed of any object relative to ourselves changing if we change our own speed. Thus, if I drop a ball while traveling on a train, the ball acquires a small downwards speed before it hits the floor. But someone watching from the platform of a station that my train speeds through will decide that the ball has a large horizontal speed as well, equal to the speed of the train. There is nothing surprising in all this, because we are used to the idea that all speeds are relative. Galileo taught us this.

But Einstein said no: light is different. The speed of light is not relative. If, on the train, I shine a flashlight forwards, then the photons travel away from me at the speed of light c , about $3 \times 10^8 \text{ m s}^{-1}$. The person on the station platform ought, if Newton and Galileo were right, to measure the speed of the photons to be larger, equal to c plus the speed of the train. But in fact, said Einstein, that person will measure the speed of the *same photons* to be exactly c relative to the platform as well. *They do not gain anything from the speed of the train!*

At first this seems impossible, since it seems to conflict with everyday experience. But we must be cautious here, and not try to shape Nature to our own preconceptions. Our experience of speeds and how they change is entirely confined

In this section: Einstein’s theory of special relativity is based on two fundamental principles. One is Galileo’s relativity principle. The other was introduced by Einstein: the speed of light will be the same to all experimenters, no matter what their state of motion. This radical departure from the way all speeds had previously been assumed to behave gives special relativity all its surprising and hard-to-accept results. Despite the difficulty we may have in accepting it, it is well verified by experiment. This is how light behaves.

►Here we refer only to the speed of light in vacuum, i.e. to the speed of free photons. When light travels through a material, like glass, it moves more slowly. Although physicists often say that the speed of light in glass is slower than the speed in vacuum, this is shorthand for what is really going on. In glass, what travels is a complicated interaction between the electric and magnetic fields of the atoms. The interaction begins when light is absorbed on one side of the glass pane, and results in light being emitted at the other side, but what is inside the glass is more complicated than just light. The fact that this interaction moves through the glass at a slower speed does not contradict Einstein’s second postulate.

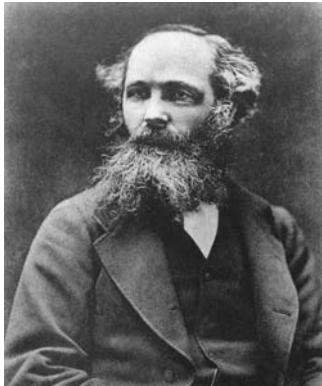


Figure 15.1. James Clerk Maxwell was one of the giants of nineteenth century physics. By unifying the apparently separate phenomena of magnetism and electricity, Maxwell opened the modern age: telecommunication and power generation technology require the cooperation of magnetism and electricity. On a more subtle level, his electromagnetism was the first unified field theory, and the first of what physicists now call “gauge theories”. This set the direction for all of fundamental physics, a direction still followed today. If this wasn’t accomplishment enough, he also made fundamental contributions to statistical mechanics, thermodynamics, and the theory of colors. He died at the young age of 48, a few months after Einstein was born. Reprinted with permission of American Institute of Physics (Emilio Segrè Archives).

to slow speeds: a train going at 100 mph (160 kph) is only moving at $1.5 \times 10^{-7}c$. The ball that falls from my hand in the train reaches the floor at a speed of only 6 m s^{-1} , or $2 \times 10^{-8}c$. We have no direct experience of combining large speeds: we might indeed shine light inside a train, but our brains and nerves are not quick enough to sense how fast the light travels down the train, or whether it is slower or faster when viewed from the platform.

Instead of relying on our low-speed experience, Einstein worked out mathematically the consequences of his two guiding principles. He found that two speeds (that of the train and that of the projectile, in our example) combine by a more complicated rule than Galileo’s. For small speeds, this rule gives the usual addition of speeds in the way that we expect, to an excellent approximation, so it is consistent with our own experience. But, for large speeds, Einstein’s rule tells us that the speeds combine only partially; in the extreme case, when one of the speeds is the speed of light, then two speeds always combine to give the speed of light exactly again, regardless of what the other speed is.

This rule, which we write down in Investigation 15.1, undermines our objection that Einstein’s prediction is contrary to our experience. Instead, his prediction is fully consistent with our experience, which deals only with the way speeds combine for small speeds. Only for large speeds, where we have no experience, does the law begin to deviate from our expectations.

Einstein had a good reason for introducing this second guiding principle, even though it seemed to contradict experience. The problem that he was trying to solve had been around since the Scottish physicist James Clerk Maxwell (1831–1879) had shown in the middle of the 1800s that electricity and magnetism are really two special cases of the general force called electromagnetism. By unifying electricity and magnetism in this manner, Maxwell had been able to show that the electromagnetic field should have waves, and he could even make a numerical prediction for the wave speed. This was close enough to the measured speed of light for physicists to realize that Maxwell’s equations had explained what light was.

However, this result had a puzzling side, because the laws said that light would travel with this speed *regardless* of the speed of the system that emitted the light. The speed of the system would affect the frequency of the light through the Doppler shift (Chapter 2), but not the wave speed. Light therefore does not behave like a projectile, something thrown out by its source, whose speed depends on the speed of the source. Instead, light behaves like a wave in a medium, having a speed that depends on the medium carrying the wave but not on the speed of the source.

We will see in the next section that most physicists of Maxwell’s time interpreted this to mean that there was a material substance that carried light vibrations at a fixed speed, regardless of the speed of the source of the light. They called this medium the ether. But nobody could find any direct evidence for the ether, and Einstein therefore decided to explore the alternative: if the laws of electromagnetism said that the speed of light was always the same number, then maybe we should just accept that as a law of physics itself. Maybe there was not an ether for it to have a speed in; maybe it just had this speed in all circumstances.

Physicists have a shorthand word for Einstein’s second guiding principle, that the speed of light is the same to all experimenters. They say that the speed of light is *invariant* under a change in the speed of the apparatus that measures it. The invariance of the speed of light is not something we can decide just by thinking about it. It is a matter for experiment. Only experiment can tell us whether to follow Einstein in his second hypothesis. And experiments on special relativity are effectively performed every day: from nuclear power generators to giant particle

▷We met the word invariant in Chapter 6, where we used it to describe situations that do not change with time or position. Here we discuss changes of experimenter.

Investigation 15.1. (Much) Faster than a speeding bullet ...

Consider two experimenters, one at rest and the other moving with speed u in the x -direction. Let the moving experimenter release a projectile moving with speed v relative to him, again along the x -axis. What is the speed V of the projectile relative to the experimenter at rest?

To Newton and Galileo, the answer was obvious:

$$V = u + v. \quad (15.2)$$

The speed of the moving experimenter adds to the speed of the projectile to give the total speed relative to the experimenter who is at rest.

But this is not consistent with the principle that the speed of light is the same to every experimenter. In particular, if $v = c$, then relativity insists that we must get $V = c$, since both speeds are the speed of a photon. Equation 15.2 does not give this result.

In other words, in special relativity speeds don't add. Physicists use the word **compose**: the speeds u and v compose to get V . The Einstein velocity-composition law is

$$V = \frac{u + v}{1 + uv/c^2}. \quad (15.3)$$

Let us try a few special cases to see what this law implies. Suppose, as before, that $v = c$. Then the numerator is $u + c$ and the

denominator is $1 + u/c = (u + c)/c$. When we divide the fraction, we get $V = c$, just as required. So this law is consistent with the principle that the speed of light is the same to all experimenters.

Another important special case of the Einstein law is when the speeds u and v are both very small compared to c . Then the fraction uv/c^2 in the denominator, which is a pure dimensionless number, is very small, and can be neglected compared with the one in the denominator. Then the Einstein law reduces to the Galilean law, Equation 15.2. This is an important consistency check: the predictions of relativity must reduce to those of Newtonian mechanics when speeds are very small compared to light.

This also explains why our intuition, based on everyday experience, leads us to expect that the Galilean law is right. Our senses and our everyday measuring devices can only deal with very small speeds, and for these the simple addition law is fine. Although we can see light, we can't sense its speed, so we don't have experience with seeing how the speed of light changed when we changed our own state of motion. If we had such sharp senses, then Galileo would have written down the full theory of special relativity from the start!

Further properties of Equation 15.3 are explored in the exercises below.

Exercise 15.1.1: More photon velocities

Let $v = -c$ in Equation 15.3, corresponding to a photon moving backwards relative to the one we tested above. Show that again $V = -c$: the speed of the photon does not depend on the observer.

Exercise 15.1.2: Computing the graph

In Figure 15.2 on the following page we plot the composition law Equation 15.3 for the special case $u = 0.4c$. Compute V/c for the set of values $v = \{0.1c, 0.4c, 0.9c\}$. Compare them with the points plotted on the curve in the right-hand panel of the figure.

Exercise 15.1.3: How fast is relativistic?

If both u and v are $0.1c$, what is the fractional error in using the Galilean addition law? [The fractional error is the difference between the Einstein and Galilean results (the error), divided by the Einstein result (the correct answer).] If $u = v = 0.3$, what is the fractional error? Suppose V can be measured to an accuracy of $\pm 5\%$. What is the largest speed (again assuming $u = v$) for which one can use the Galilean formula and make errors too small to be measured?

Exercise 15.1.4: Zero is still zero

The Einstein composition law still has some features that we expect from everyday life (and logical consistency). Show that, if the projectile remains at rest with respect to the moving experimenter (so $v = 0$), then its speed relative to the experimenter at rest is the same as the speed of the moving experimenter, $V = u$. Show further that if the moving experimenter shoots the projectile backwards with a speed of $v = -u$, then it will be at rest with respect to the resting experimenter ($V = 0$).

accelerators to the GPS navigational satellite system that we mentioned in Chapter 2, many of today's high-technology devices would not function correctly if special relativity were wrong. Special relativity is one of the best-tested aspects of all of fundamental physics.

We should note here that physicists tend to use other words for what we call **experimenters** in this book. When you read other books on relativity you may read about **observers** or **frames**. Observers are the same as experimenters, and we will sometimes use the two terms interchangeably. A frame is the coordinate system used by the observer or experimenter to locate events in space and time, and we will have much to say about coordinates in these chapters on relativity. But we will not use the word "frame" in this book, except in Chapter 24, where we use it to describe a coordinate system spanning the whole Universe, something that is not really easy to envision as a single experimenter or observer. Otherwise we avoid the term "frame", because it is impersonal, and it might lead you to think that the results of special relativity are somehow to do with bad definitions of coordinates. The words "experimenters" and "observers" are more appropriate, because they should make you think of careful scientists who make measurements with the best techniques,

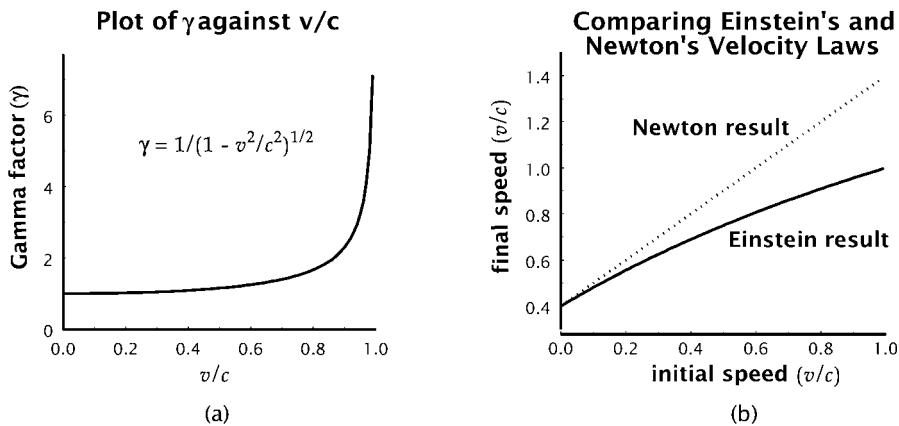


Figure 15.2. The two figures show how much the predictions of Einstein differ from those of Galileo and Newton. (a) The left-hand panel shows the Einstein γ -factor, $\gamma = 1/(1 - v^2/c^2)^{1/2}$. For a given speed v (shown as a fraction of c on the horizontal scale), the height of the graph gives the factor by which time stretches, masses increase, and lengths contract at that speed. This factor is always bigger than one, and it gets infinitely large as v approaches c . (b) The right-hand panel compares one example of the Einstein law for combining speeds with that which Newton and Galileo would have expected (and which our low-velocity intuition leads us to expect, too). Suppose a rocket ship is moving at the fixed speed $0.4c$ relative to us, and suppose it fires a particle forwards with speed v relative to the rocket. Then the two curves show the two predictions of the speed of this particle relative to us (vertical scale) as a function of its speed relative to the rocket (horizontal scale). For the Einstein law (solid curve), the final speed never gets bigger than c , and only reaches c when the particle's speed relative to the rocket is c . This means that a photon fired from the rocket will have speed c relative to the rocket and to us, in accordance with one of the founding principles of special relativity. By contrast, the dotted line shows the speed that Newton and Galileo would have expected, namely $0.4c + v$, which can exceed c . Simple as this formula is, Nature follows the Einstein law instead.

and who really do measure all the unexpected results of special relativity.

The Michelson–Morley experiment: light presents a puzzle

In this section: the experimental foundations of special relativity go back to the Michelson–Morley experiment. It gives direct evidence that the speed of light is an invariant. Until Einstein, physicists could not make sense of the result of this experiment, and most of them simply ignored it. It is not clear whether Einstein himself realized its importance until after he was led to his postulates by examining Maxwell's equations.

►Interferometry is a classic example of a phenomenon seen repeatedly over the centuries, where a key technology is invented by a scientist simply as a tool for the investigation of a deep question in "pure" science. Interferometry is used today in industrial machining, geology, pollution control, astronomy, and countless other areas.

In fact, the earliest experiment that supported Einstein's guiding principle on the speed of light was actually done well before Einstein's 1905 paper on special relativity. It is the famous Michelson–Morley experiment, performed in the 1880s. It was designed to test how the speed of light changed in certain circumstances. When it found that there was no measurable change, it became one of the puzzles that physicists of the day could not understand. Only after Einstein's work did scientists accept that the Michelson–Morley experiment had been telling them all along about the invariance of the speed of light.

It is worthwhile spending a little of our time looking at the Michelson–Morley experiment. Not only will it show us that Nature really does adhere to Einstein's second guiding principle, but it is also a chance to look at a remarkable invention, the Michelson **interferometer**. This instrument, which the American scientist Albert A Michelson (1852–1931) devised in order to measure changes in the speed of light, has developed into one of the most important high-precision measuring instruments of modern science and technology. And it is being developed further today to detect one of the most significant predictions of general relativity: gravitational waves. We will look closely at this instrument in the next section. This discussion will prepare us for studying gravitational waves and their detection, in Chapter 22.

Michelson invented the interferometer to perform his first experiment on light in 1881. This produced such an unexpected result that he repeated it with greater precision with his American collaborator Edward W Morley (1838–1923) in 1887. The aim was to show that light had a different speed relative to the laboratory when it traveled in the direction of the Earth's motion than when it moved in a perpendicular direction.

In fact, Michelson expected it to be *slower* along the Earth's motion than across

it. From Maxwell's theory, physicists knew that light was a wave, and that it was just a short-wavelength version of radio or other electromagnetic waves. For most physicists, if something was a wave then it had to be a vibration in *something*, and they called this medium the *ether*. This was a hypothetical substance whose vibrations were light waves.

The problem with the ether was that it had to be everywhere, in order to carry light to us from the distant stars, and yet there was no independent evidence for it. For example, if the planets were moving through the ether on their orbits around the Sun, why did it not slow them down? Why did the planets follow Newton's laws so exactly? To get around this, physicists had to assume that the ether was frictionless, unlike any other substance known. Many physicists were uncomfortable with such implausible properties, and in fact it was Einstein's own discomfort that led him to throw out the idea of the ether and embrace the invariance of the speed of light. But in the 1880s physicists were not ready for this. Instead, they felt that they had to find direct evidence for the ether.

Michelson hit on a way to do this. He reasoned that, if light had a fixed speed relative to this medium, then as the Earth traveled through the medium, the speed of light relative to the Earth would be slower in the direction of the Earth's motion than perpendicular to it. So Michelson expected that by comparing the speed of light in the direction of motion of the Earth as it orbits the Sun with the speed of light in the perpendicular direction, he would be able to measure the speed of the Earth relative to the ether, and thereby give a strong demonstration that the ether was really there.

Instead, what he found must have been very frustrating to him, at least at first: he could detect no difference between the speeds, so it seemed that the Earth was always at rest with respect to this ether, regardless of its motion. The ether was undetectable in this experiment.

It is interesting to look at the way physicists in 1887 reacted to this experiment. The cleanest thing to do, after this, would have been to abandon the ether idea: if it is not measurable, maybe it does not exist. This would have required a radical reformulation of physics, however, and no-one was able to do this until Einstein, 18 years later. In fact, most physicists found the experiment difficult to incorporate into their thinking. Many recognized the importance of the experiment, but could not fit it into the rest of what they knew about physics. Some went to extremes to defend the ether, such as postulating that the ether, far from being frictionless, was dragged around by the Earth; but such "fixes" raised other problems and were clearly contrived. Other physicists simply put the problem aside as being too difficult, and they worked on something else, where they knew they could make progress.

The most important attempts actually to find a plausible way to explain the Michelson–Morley result were by the Dutch physicist Hendrik A Lorentz (1853–1928) and the Irish physicist George F Fitzgerald (1851–1901). They pointed out, independently of each other, that if the interferometer actually *contracted* by a certain amount in the direction of its motion, then this would compensate the expected smaller speed of light in that direction and allow the photon going this way to return at exactly the same time as the photon moving across the motion. The mathematical content of their work was elaborated by French physicist Henri Poincaré (1854–1912).

We shall see below that a length contraction is indeed a prediction of Einstein's theory, and we call it today the **Lorentz–Fitzgerald contraction**. But the contraction in Einstein's theory does not take place in the same circumstances as Lorentz and Fitzgerald predicted, and Einstein's explanation of the Michelson–Morley ex-

>When Einstein derived the Lorentz–Fitzgerald contraction from his two fundamental principles, Lorentz and Poincaré were among the first physicists to recognize the significance of his new approach, to applaud the genius of this young patent clerk in Switzerland, and to help open the doors of the academic world to him.

In this section: the instrument that Michelson invented to perform the Michelson–Morley experiment is the *interferometer*. Today this forms the basis of the instruments being built to detect gravitational waves, a prediction of Einstein's general theory of relativity. We explain here how the instrument works.

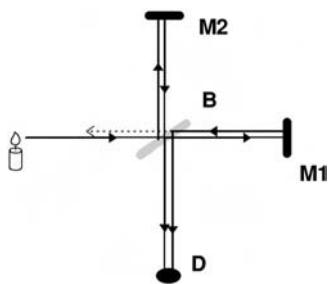


Figure 15.3. A sketch to show the principle of the Michelson interferometer

periment is very different from that of Lorentz and Fitzgerald. We will return to this difference below.

Michelson's interferometer: the relativity instrument

Michelson's interferometer, shown in Figure 15.3, is basically a device for measuring the tiny differences in the time it takes light to travel along two perpendicular paths. This can't be done using a stopwatch, because light travels too fast. Instead, the device only compares the two times, measuring the *difference* in light-travel-time along the two directions, but not telling us what the travel time along either direction actually is. To see if light travels at different speeds in two different directions, this is all one needs.

Here is the simple idea behind the device. Suppose it can be arranged that two photons leave the light source (the candle) at exactly the same time, so they reach the point B at the same time. Suppose further that one photon then travels to mirror M1 and the other to mirror M2. If the lengths B–M1 and B–M2 are the same, then when reflected, they will return to B at the same time if and only if they travel at the same speed in the two directions. If one of the speeds is larger, then the photon traveling that way will return before the other one. If one can measure the *difference* in arrival times, one can measure the difference in the speeds. In Investigation 15.2 we look at how, by using the interference of light in the two arms, the interferometer can measure tiny differences in arrival times.

In the Michelson–Morley experiment, the instrument is carried through space by the motion of the Earth. By aligning one "arm", say B–M1, in the direction of motion of the Earth, and the other arm (B–M2) across that motion, the experimenters expected to see a difference in speeds along the two arms. *They saw no difference.* No matter how the instrument was oriented, no matter what direction it was carried by the Earth, the two photons both arrived back at the point B at exactly the same time. The accuracy of their experiment was good enough to see a difference even if the speed of the Earth through the ether were only 1% of the speed of the Earth around the Sun. But they saw nothing.

Following Einstein, we now interpret Michelson and Morley's result as a direct demonstration that the speed of light does not depend on the speed of the instrument measuring it.

It is not hard to see how an interferometer could be used for other high-precision experiments today. Since we know that the speed of light is the same in each arm, any difference in the arrival times of light after traveling in the two arms must be caused by a difference in the *lengths* of the arms. Interferometers today are used to make sensitive length measurements. We will see that this is exactly what is needed when looking for gravitational waves, which can make minute changes in the arm-lengths of a suitably constructed interferometer.

Now, this description of how the interferometer works is clearly somewhat oversimplified. The principle is correct, but there are many impractical aspects of the description. It is not practical to get just two photons to leave the light source at the same time, and it is not practical to measure the difference of their arrival times back at point B directly. It is even rather difficult to insure that the two arms B–M1 and B–M2 are exactly the same length. Readers who want to find out how Michelson actually did it will find a more realistic description in Investigation 15.2. But readers who do not consult this investigation will not miss anything essential. In particular, our discussion of how modern astronomers expect to use Michelson's interferometer to detect gravitational waves will require only the ideas just described.

Investigation 15.2. How an interferometer works, and why it got its name

In the Michelson experiment, the light source is a continuous beam of light, not just an emitter of two photons. The difference in arrival times at the detector D is detected by allowing the two beams to *interfere* with one another.

Interference should be a familiar phenomenon to anyone who has watched water waves in a harbor or in a bathtub. When waves pass through each other, there are places where the height of the water goes up and other places where it is cancelled out. The high places are places where the peaks of the two individual waves coincide. The low places are places where a peak of one wave coincides with a trough of the other. If the two waves have the same wavelength, then the pattern of peaks and valleys where they interfere can stay in one place for a relatively long time.

Light behaves in the same way, as an oscillating electromagnetic wave. When two light beams interfere, they make a pattern of light and dark spots or stripes. These are called fringes. In order to get a good interference pattern, it helps to use light of a single wavelength. Broad-band light can be made to interfere in one place, but if the two beams have similar wavelengths then they will make an interference pattern over a wider region. This makes it easier to set up the interferometer and to live with small differences in arm-length. Michelson filtered his light to a narrow band of colors; today scientists typically use monochromatic (single-color) lasers as the light source.

Consider, therefore, a single-color beam of light leaving the light source (the candle in the diagram) and reaching the beam splitter B, drawn as a gray diagonal element in the diagram. This is a half-silvered mirror, which means that half the light goes through it and

half is reflected. These two beams of light leave the beam splitter with their oscillation peaks and valleys locked in step together, since they came from the same original beam.

After reflecting from the mirrors M1 and M2, they arrive back at the beam splitter. Both beams are split again, with half the wave reflecting and half transmitting. Here is where the interference phenomenon takes place. If the arms are exactly the same length, and the speed of the light in the arms is the same, then the beam coming from M1 that goes through the beam splitter back towards the light source, and the beam from M2 that is reflected by the beam splitter toward the light source, will be exactly in phase with each other, and they will add together to make a strong beam leaving the interferometer in this direction. At the same time, the light beams that head toward the detector D will be exactly out of phase with each other, and they will nullify one another so that *no* light goes to the detector. One could make other configurations: if the two arms differ, for example, by one-quarter of a wavelength, then the light will all go to D, with nothing going back toward the source.

What Michelson expected was that, during the course of the experiment, as the Earth turned and changed the direction of the arms of his interferometer, the speed of light in one arm would change relative to the other, and this would change the interference arrangement. So he looked for changes in the amount of light falling on the detector D that had a period of 12 h, the time it takes the interferometer to rotate to an equivalent configuration. He could detect changes induced by differences of speed between the two arms as small as 1% of the expected difference in speed (the speed of the Earth around the Sun), but he saw none.

Special relativity: general consequences

The Michelson–Morley experiment shows us that Einstein's principle that the speed of light should be the same for all experimenters is correct, even though it is radically different from our expectations. A theory founded on such a radical idea is bound to have consequences that are equally radical. In this section I shall list the important consequences of special relativity that we will need to know about in order to go on and study black holes and cosmology.

The list in this section will be brief but comprehensive. It will cover all the effects of special relativity that we will need in our discussions of general relativity and its astronomical implications. Readers who seek a deeper understanding of special relativity itself will find a section on each of the following points in the next chapter, Chapter 16. These sections will give more derivations where necessary, discuss the interrelations between these points, treat the experimental support for various effects, and dispose of worries that there might be internal contradictions (paradoxes) within special relativity.

1. *Nothing can travel faster than light.* No matter what forces one uses to accelerate a particle (or a rocket) to higher speeds, the object will never reach the speed of light. This comes from Einstein's formula for the combination of speeds, as illustrated in Figure 15.2 on page 184.
2. *Light cannot be made to stand still.* Obviously, since light has the constant speed c , it cannot be brought to rest. This principle applies to light that is free to move in empty space. In many circumstances, light interacts with other things: it bounces off a mirror or travels through a piece of glass. In these circumstances the "light" wave can travel at different speeds (or even come to rest instantaneously as it is being reflected from the mirror). But in this case the speed refers, not to pure light, but to a property of the interaction of light with other matter. In a vacuum, light travels at one speed only.

In this section: we list the most important consequences of the principles of special relativity, along with brief explanations and key formulas. Each consequence is treated in more detail in the next chapter. The list includes:

- nothing can travel faster than light;
- light cannot stand still;
- time slows for moving bodies;
- moving objects contract in length;
- simultaneity depends on the observer;
- mass depends on speed;
- energy and mass are equivalent;
- photons have zero rest-mass;
- the Doppler effect is changed.

▷ The speed of light is the limit on all speeds.

▷ Anything that travels at the speed of light cannot be made to come to rest.

▷ Time runs slower for moving bodies. It stands still for light. This is called time dilation.

3. *Clocks run slower when they move.* It is not possible to lose Galileo's simple formula for adding speeds together without also losing his simple notions of space and time. In order that the speed of light should be the same to two different experimenters, something unexpected must happen to the way they measure time and space, since a speed is simply the ratio of a distance to a time. What happens to time is that it slows down at high speed. For example, if an unstable elementary particle decays in a time t as measured by an experimenter at rest with respect to the particle, then it will decay in a time

$$t' = \frac{t}{\sqrt{1 - v^2/c^2}},$$

as measured by an experimenter moving past the particle at speed v . Here we meet for the first time an expression that is so important in equations in special relativity that it is given its own symbol, γ :

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}}. \quad (15.4)$$

In the left-hand panel of Figure 15.2 on page 184, I have drawn γ as a function of v/c . For a speed v that is small compared to that of light, this is nearly equal to 1, and the two times differ by very little. Only when v becomes close to c do relativistic effects become important. The slowing of time is called **time dilation**. Notice that as v approaches c , t' gets longer and longer. In other words, for a photon (going at speed $v = c$), time stands still. Photons do not age. For this reason, they also do not decay: the only way a photon can change is to interact with something outside it. We referred to this in Chapter 11, to show that if neutrinos change their type as they move, then they must have mass.

▷ The size of a moving object contracts along its motion. This is called the Lorentz–Fitzgerald contraction.

4. *The length of an object contracts along the direction of its motion.* Not only time is affected: democratically, lengths also depend on speed. If an experimenter at rest with respect to, say, a length of pipe, measures its length to be L , then an experimenter moving past with speed v will measure the length L' to be shorter, by the same factor as time got longer:

$$L' = L\sqrt{1 - v^2/c^2} = L/\gamma. \quad (15.5)$$

This is the *Lorentz–Fitzgerald contraction*, and is the same formula originally written down by Lorentz and Fitzgerald as they tried to explain what happened in the Michelson–Morley experiment.

In special relativity, this is really just a counterpart to the time dilation: the two are two faces of a single coin. To see this, consider a rocket ship crossing the Galaxy at such a speed that $\gamma = 10^{12}$. Now, the Galaxy has a diameter of about 100 000 light-years, and the rocket is traveling at very close to the speed c , so the clocks on Earth tick about 100 000 years while the rocket makes the trip. But the clocks on the rocket tick at only $1/10^{12}$ of this time, which is 3 s! However, on the rocket the astronauts do not feel any different: by the principle of relativity, they consider themselves to be at rest, and every clock on board is ticking at its normal rate. How, then, can they cross the Galaxy in only three of their seconds?

Length contraction is the answer: the Galaxy has, as they measure it, a diameter of 100 000 light-years divided by 10^{12} , or only 10^9 m. Since the Galaxy

is flashing past them at nearly the speed of light, it takes only 3 s to go completely past. So length contraction can be derived from time dilation, and vice versa.

This illustrates a fundamental aspect of relativity, that different observers (in this case the astronauts and an observer at rest in the Galaxy) must always make the same prediction for the outcome of an experiment (in this case the number of seconds ticked on an astronaut's wristwatch as he crosses the Galaxy), but they may explain the outcome of the experiment in different ways (time dilation or length contraction).

5. *There is no universal definition of time and simultaneity.* If two events that happen in different places are measured to occur at the same time by one experimenter, they may not be measured to occur at the same time by other experimenters that are moving with respect to the first. We call this the **loss of simultaneity**. To Newton and Galileo, *before* and *after* had invariant meanings: everyone would agree that event A happened before event B. This seemed only logical, since event A might have been part of the cause of event B, and it would be contradictory if some else determined that B occurred first. In relativity, this logic only requires that the notion of before and after is required to apply only to events that can influence one another. Thus, if A could cause B, then everyone must agree that A was earlier. But A can only cause B if light can travel from A to B: no influences travel faster than light. Therefore, if B is too far away for light to get there from A by the time B happens, then there can be no cause-and-effect relation and there is no logical need for different experimenters to agree on which one occurred first.

Events that occur at the same time in different places as measured by one experimenter are exactly of this type: neither can be the cause of the other. So it happens that relativity does not give them a unique order: to one experimenter they are simultaneous, to another A may occur first, and to a third B may occur first. But all three experimenters will agree that light cannot make it from one to the other, so they can have no causal effect on each other.

On the other hand, if light *can* travel from A to B, then all experimenters will also agree on this, and all will place B later than A (though by differing amounts of time, depending on the time dilation effect). So relativity preserves a notion of before and after, of *future* and *past*, but it does not apply that relation to all possible pairs of events.

This means that it is not possible to maintain Newton's idea of a three-dimensional absolute space, for which time is just a parameter: in Newton's world everyone would agree on what space looked like at a given time. In Einstein's world, there is just **spacetime**, the four-dimensional continuum of all events that occur anywhere at any time. Notice that **events** are the "points" of spacetime: an event is something that occurs in a particular location at a particular time, so it is a "dot" in spacetime. One experimenter will group a particular set of events into 3D space at a particular time, but a different experimenter could equally validly decide that a very different set of events constituted space at that particular time.

Two events that cannot have a cause-and-effect relationship with one another are said to have a **spacelike** separation in spacetime. Two events that can be connected by something traveling at less than the speed of light are said to

▷ Simultaneity is not something that all experimenters will agree on. This disagreement is closely tied to the time dilation and length contraction.

have a **timelike** separation in spacetime. Two events that can be connected by a single photon are said to have a **lightlike** separation.

Relativity mixes notions of time and space. If we change point of view from one experimenter to another in relative motion, then there is a transformation in how we distinguish space from time, in how we reckon the passage of time, and in how we measure distances. This whole change in point of view is called the **Lorentz–Fitzgerald transformation**. It has a mathematical expression, but we need not deal with that. The main thing is that any one experimenter can use his or her own conventions on time and space consistently, but there is nothing absolute about them. Another experimenter's conventions will do just as well, even though they are different. This mixing of time and space will be discussed thoroughly in Chapter 17, where it will form the basis of our study of general relativity.

▷ As an object moves faster, its inertial mass increases, so it is harder to accelerate it. This enforces the speed of light as a limiting speed: as the object gets closer to the speed of light, its mass increases without bound.

6. *The mass of an object increases with its speed.* We noted above that no force, no matter how strong, could accelerate a particle to the speed of light. Does this mean that Newton's second law, $F = ma$, is wrong? After all, if I take F to be large enough, I should be able to make the acceleration large enough to beat the speed limit of c . No: relativity has a way out of this potential contradiction. The mass m in Newton's equation gets unboundedly large as the particle gets near to the speed of light, again by the ubiquitous γ factor:

$$m = \gamma m_0, \quad (15.6)$$

where m_0 is the mass that an experimenter at rest with respect to the particle would measure. Physicists call m_0 the particle's **rest-mass**. In fact, they define the rest-mass of any object to be its inertial mass when it is at rest. If the object is complex, like a gas with lots of random internal motions, then the rest-mass is the mass when the average momentum of all the particles is zero.

7. *Energy is equivalent to mass.* Here we meet the most famous equation associated with Einstein:

$$E = mc^2.$$

To see what it means, and why it makes sense, consider again what happens if we try to accelerate a particle up to the speed of light. We keep applying an immense force to it, but since its mass is increasing rapidly, its speed hardly changes. Nevertheless, the force is doing work, by Equation 6.20 on page 62, and we have to keep supplying energy to keep the force going. What is happening to this energy? In conventional Newtonian language, we would at least expect the kinetic energy of the particle to be increasing. This was introduced in Equation 6.8 on page 54. Einstein showed that the energy accounting comes out right if the total energy of the particle is just its total mass times c^2 . This includes its kinetic energy and, of course, a new energy: the rest-mass energy m_0c^2 that the particle has even when it is at rest.

This new concept had many important implications for physics. For one, it meant that energy has inertia: the more energy one puts into a system, the more mass it has and therefore the harder it is to accelerate. For another, it became possible to imagine the conversion going the other way: reducing the mass of an object and releasing the corresponding energy in another form. This is the implication that is most familiar to us: nuclear reactors and nuclear explosives work in this way. But it is important to understand that this conversion happens in everyday life, too, although on a scale that we don't notice.

▷ Einstein's most famous equation expresses the *equivalence* of energy and mass, not just the ability to convert between them. The kinetic energy of a moving body accounts for its increased inertial mass. Any object that gains energy, say from heat, is also harder to accelerate because it has more mass.

If an automobile has a rest-mass $m_0 = 1000 \text{ kg}$ (its mass when it is standing still), then when it is moving at speed $v = 100 \text{ km hr}^{-1}$ (about 60 mph), so that it has $v/c = 9 \times 10^{-8}$, its total mass is larger by about $4 \times 10^{-12} \text{ kg}$. (See Exercise 15.3.2 on the next page for the details of this calculation.) If two such cars collide head-on and come to rest, then the rest-mass of the wreck is larger than the sum of the two original rest-masses by twice this amount, or about 8 picograms (less, of course, the mass of the hubcaps and other pieces that roll away, the mass-equivalent of the sound energy radiated by the collision, the mass left in tire skid-marks on the road, and so on). This 8 pg of mass takes the form mainly of extra chemical energy in the deformed structures of the cars. It is a real mass: it would show up in a precision weighing of the wreck, and it would contribute to the inertia of the wreck if the rescue vehicles try to push the mess off the road.

8. Photons have zero rest-mass; their momentum is proportional to their energy.

When a particle is accelerated to nearly the speed of light, its mass increases without bound. How, then, do photons get to the speed of light with finite energy? The only consistent answer is that their rest-mass should be zero. This is not really a well-defined notion, since photons cannot be brought to rest in order to measure their rest-mass. It is really only a convenient way of speaking about photons to explain why they do not fit into the rest of mechanics. But it is also a new perspective that allows us to speculate that perhaps there are other particles that have zero rest-mass as well. They, too, would travel at the speed of light. From this point of view, the “speed of light” is more fundamental than light itself: it is the speed of all zero-rest-mass particles. The neutrino (see Chapter 11) was at first thought to be massless and to travel at speed c , although observations today suggest that it has a very small mass. When gravity is turned into a quantum theory, some physicists expect that there may be a particle associated with gravitational waves, called the **graviton**, which will also be massless. But, as we shall discuss in Chapter 27, it may be very different from the photon.

We have seen that the momentum carried by a photon is important in astronomy. The momentum of any particle is its mass times its speed. In relativity, the mass is its total mass m . This is equivalent to its total energy E divided by c^2 . For a photon, whose speed is always c , the momentum is therefore

$$\text{photon momentum} = (E/c^2) \times c = E/c.$$

9. The Doppler redshift formula changes slightly. The changed notion of time in special relativity leads to a simple modification of the formula for the redshift of a photon. Remember how we derived the formula, by a visual method using Figure 2.3 on page 15. We counted the number of wave crests that passed by a moving wave-crest counter, and compared that with the number that passed one at rest. The number of crests passing per unit time is the frequency of the wave. Now we have to take into account that the moving counter’s clock is running a bit more slowly than the one at rest. So if the counter at rest counts a number N crests in a time t , the moving counter counts a number $N' = N(1 - v/c)$ crests (from Figure 2.3 on page 15) in a time $t' = t/\gamma$ (Einstein’s time dilation). When we divide the number of crests by the time, the counter at rest measures a frequency $f = N/t$, while the moving

►Traveling at the speed of light, photons are special. They have no rest-mass and they carry a momentum that is proportional to their energy, a very different relationship from the one that governs non-relativistic particles.

►Because of the time dilation effect, even velocities across the line-of-sight to a body will slow time down and therefore change the apparent frequency of light it emits. So the Doppler formula must be modified to take account of time dilation.

Investigation 15.3. Relativity at small speeds: making Galileo happy

The formulas of special relativity look rather complicated when one first meets them, with all those factors of v^2/c^2 and $(1 - v^2/c^2)^{-1/2}$. All these factors are necessary to deal with phenomena at or near the speed of light. But when speeds are small, we must expect to get the Galilean and Newtonian results as well. We saw how this worked with the velocity composition law in Investigation 15.1 on page 183. We want to look at it more systematically here, since there are so many formulas with these factors in them.

One can use the binomial theorem Equation 5.1 on page 43 to show that, for small speeds v ,

$$\left[1 - \left(\frac{v}{c}\right)^2\right]^{-1/2} \approx 1 + \frac{1}{2} \frac{v^2}{c^2}. \quad (15.8)$$

The term $v^2/2c^2$ is an estimate of the size of the relativistic correction to a Galilean or Newtonian formula. For example, the mass of a particle with rest-mass m is larger than the rest-mass by

$$\Delta m = m \left[1 - \left(\frac{v}{c}\right)^2\right]^{-1/2} - m \approx \frac{1}{2} \frac{v^2}{c^2} m.$$

The equivalent excess energy is $\Delta mc^2 \approx \frac{1}{2}mv^2$. This is what is called the kinetic energy in non-relativistic physics. We see that it is a low-

velocity approximation to the correct value of the energy associated with the motion of the particle.

Similarly, in the Lorentz-Fitzgerald contraction, the change of length is

$$\Delta L = L \left[1 - \left(\frac{v}{c}\right)^2\right]^{1/2} - L \approx -\frac{1}{2} \frac{v^2}{c^2} L.$$

If the speed is 1% of the speed of light, for example, then the contraction is only 0.005% of the original length.

If we consider the volume of a rectangular box which is made to move parallel to one of its sides, then it will contract along that length and not along the other two, so its volume will decrease in direct proportion to the length of the contracting side. The above formula then has the consequence that for small speeds the change in volume is approximately

$$\Delta V = -\frac{1}{2} \frac{v^2}{c^2} V. \quad (15.9)$$

We will need to use this below when we discuss the dynamics of moving fluids.

Exercise 15.3.1: Slow-velocity expansion

Use the binomial expansion Equation 5.1 on page 43 to show that the expansion of $(1 - v^2/c^2)^{1/2}$ for small v/c is

$$\left[1 - \left(\frac{v}{c}\right)^2\right]^{1/2} = 1 - \frac{1}{2} \left(\frac{v}{c}\right)^2 + \dots$$

Exercise 15.3.2: How much mass is in kinetic energy?

Consider the example given in the text, of an automobile with a rest-mass of 1000 kg. Show that its kinetic energy at a speed of 100 km hr⁻¹ has a mass equivalent of 4 pg (4 picograms, or 4×10^{-12} g).

counter gets $f' = N'/t'$, which works out to be

$$f' = (1 - v/c)\gamma f = \frac{1 - v/c}{\sqrt{1 - v^2/c^2}} f. \quad (15.7)$$

This is the formula when the moving counter is going away from the source of light, as seen by the counter at rest. This produces a decrease in the frequency, or a redshift. If the moving counter is approaching the source of light, there is a blueshift, an increase in the frequency, because the sign of v changes and the numerator in this equation is then bigger than one. Because the denominator is always smaller than one, the redshift and blueshift are larger than one gets from the non-relativistic Doppler formula. Notice that there is even a Doppler shift if the moving counter is moving *perpendicular* to the direction to the source of light. In this case, the non-relativistic Doppler shift is zero, because the motion of the counter does not add or subtract any wave crests from the number counted by a counter at rest. But there is still the time dilation, which reduces the amount of time that the moving counter measures while it counts the crests. This produces a blueshift in relativity where there is none in the Newtonian Doppler formula. This is called the transverse Doppler shift.

The extra inertia of pressure

In this section: we single out an unexpected consequence of special relativity: the more pressure a gas has, the harder it is to accelerate.

lated bodies. In our tour of the Universe we have studied gases in order to understand stars, and at the end of our tour we will again need to understand gases in order to understand the Universe as a whole. In this section we will discover, perhaps rather unexpectedly, that in special relativity the pressure inside a gas plays an important part in the inertia of the gas: the more pressure the gas has, the more difficult it is to accelerate.

While at first sight this might seem like a mere curiosity, it has very far-reaching consequences. When we study neutron stars in Chapter 20 we will find that this effect makes the neutron gas “weigh” more, and this in turn forces the star to have a higher pressure, which only makes it weigh even more, and so on. This pressure-feedback effect eventually makes it impossible for the star to support itself: the inertia of pressure opens the door to the black hole.

The inertia of pressure can be traced to the Lorentz–Fitzgerald contraction. In Investigation 15.4 on the next page we show how to calculate the extra inertia, but even without much algebra it is not hard to see why the effect is there. Consider what happens when we accelerate a box filled with gas. We have to expend a certain amount of energy to accelerate the box, to create and maintain the force of acceleration. In Newtonian mechanics, this energy goes into the kinetic energy of the box: as its speed increases so does its kinetic energy. This happens in relativity too, of course, but in addition we have to spend some extra energy because the box contracts.

The Lorentz–Fitzgerald contraction is inevitable: the faster the box goes, the shorter it gets. But this shortening does not come for free. The box is filled with gas, and if we shorten the box we reduce the volume occupied by the gas. This compression is resisted by pressure, and the energy required to compress the gas has to come from somewhere. It can only come from the energy exerted by the applied force. This means the force has to be larger (for the same increase in speed) than it would be in Newtonian mechanics, and this in turn means that the box has a higher inertia, by an amount proportional to the pressure in the box.

In fact, the formula for the extra inertia is simple. If the box has a mass-density ρ (which, in relativity, includes the mass associated with all the different forms of energy in the gas) and pressure p , then the density of inertial mass is

$$\text{inertial mass density} = \rho + p/c^2. \quad (15.10)$$

This equation is derived in Investigation 15.4 on the following page. It is simple to use. If the box has a volume V , then the total inertial mass in the box is $(\rho + p/c^2)V$, in the sense that force F required to produce an acceleration a is just $F = [(\rho + p/c^2)V]a$. This is simply a consequence of special relativity. We will find that inertial mass density useful in our study of cosmology: it is a key to understanding Einstein’s cosmological constant and the theory of inflation.

Conclusions

The ideas we have discussed and the formulas we have derived in this chapter will lead us naturally into relativistic gravity and its consequences in later chapters. However, some readers may want a more detailed discussion of the points described here. For example, why is the time dilation not an internal contradiction in special relativity: if experimenter A measures that the moving clocks of experimenter B are going slowly, how can relativity be preserved? After all, the principle of relativity says that an experiment should not single out a preferred speed. Thus, if experimenter B measures the rate of A’s clocks, which after all are moving with respect to B, then B should find that they are going more slowly. But how can the clocks

In this section: our survey of special relativity will be sufficient for the rest of the book, but readers can find more depth in the next chapter. For general relativity, skip to Chapter 17.

Investigation 15.4. How pressure resists acceleration

We shall look at this only for slow motions, where the effects of special relativity are small. Suppose we have a box at rest that is filled with a uniform gas. We denote the volume by V , the mass density by ρ , and the pressure by p . Suppose next that we apply a small force to the box and accelerate it until it has a speed v that is small compared to c . The key question is, how much energy did we have to put in to get the gas up to speed v ? For simplicity, we will only ask about the gas, not about the container: in astronomy we usually don't have containers: one part of the gas of a star is held in place by gravity and the pressure of other parts of the gas.

Once it is at speed v , the gas in the box has acquired a kinetic energy, so one might think that the total energy that we had to add to the box in order to accelerate the gas in it would have been equal to this kinetic energy, $\frac{1}{2}mv^2 = \frac{1}{2}\rho Vv^2$, where in the second expression we have used the fact that the mass m in the box is ρV . But this is not the whole story, because the Lorentz–Fitzgerald contraction has shortened the length of the box and therefore changed its volume. Making a box smaller when it contains a fluid with pressure p requires one to do work on it, in other words to put some energy into the gas. This extra energy represents the extra inertia of the gas: it is harder to accelerate the gas because it takes work not only to accelerate the existing energy but also to compress the gas as the Lorentz–Fitzgerald contraction demands.

We only need to work out this extra energy in order to see why. The energy one has to put into it is just $-p\Delta V$, where we denote the change in volume by ΔV ; the minus sign is needed so that when the box contracts (ΔV negative) then the energy put into the box is positive. Using Equation 15.9 on page 192 to get the change in volume, we find that the extra energy we put in is

$$\frac{1}{2} \frac{v^2}{c^2} pV.$$

This energy does not just disappear; it goes into the internal energy of the gas in one form or another, depending on the details of the gas molecules. At least some of the energy goes into raising the temperature of the gas (the random kinetic energy of the molecules).

The total energy required to accelerate the gas-filled box can be written in a simple way:

$$\begin{aligned} E &= \frac{1}{2}mv^2 - p\Delta V \\ &= \frac{1}{2}\rho Vv^2 + \frac{1}{2} \frac{v^2}{c^2} pV \\ &= \frac{1}{2} \left(\rho + \frac{p}{c^2} \right) v^2 V. \end{aligned} \quad (15.11)$$

The last expression is the one we need to examine. The energy required to accelerate the box is proportional to the sum $\rho + p/c^2$. This energy comes from the work done by the force we must use to accelerate the box, so the force had to be larger than we might have expected. Put another way, for a given applied force, the box accelerates less than we would have expected by measuring its mass, since some of the energy we put in goes into the internal energy of the gas instead of the kinetic energy of the box. Scientists therefore say that the inertia of the box is larger than just its rest-mass, and in particular they call the quantity $\rho + p/c^2$ the *inertial mass density* of the gas. If we want to know how much force is required to accelerate a fluid we have to know the inertial mass density, not just the rest-mass density. This is purely a consequence of special relativity.

of B go more slowly than those of A, and at the same time the clocks of A are going more slowly than those of B? The same worry arises for the length contraction. In fact, these apparent contradictions are not real: they result from not considering carefully enough what is being measured. To allow us more space for a discussion of these deep and profound issues, and also to give readers a glimpse of the enormous body of experimental evidence that now supports special relativity, each of the points discussed in the list earlier in this chapter is given a separate section in the next chapter. The interested reader can use these sections to become much more deeply acquainted with special relativity.

This extra material is not essential for our investigations of relativistic gravity in later chapters, so readers who want to stick to the main line of development can safely leave them out and go straight to the first chapter on general relativity, Chapter 17.

Relating to Einstein: logic and experiment in relativity

Our introduction to special relativity in the last chapter covered the basics, but it may have raised more questions for you than it answered. Before reading the chapter, you may have been very happy with the simple idea that everyone would agree on the length of a car, or the time it takes for the hands on a clock to go around once. If so, you have now learned to question these assumptions, that Nature does not really behave like that. If you want to fit these ideas together into a more logical framework, and if you want to learn something about why scientists are so sure that Nature really follows the principles of special relativity, then this chapter is for you. Read on.

In the previous chapter I listed some important effects of special relativity and gave a brief description of each, such as time dilation and the equivalence of mass and energy. In most cases, I left out the derivations, the algebra that linked one result to another. In this chapter I will fill in some of these gaps. Each of the points in Chapter 15 has its own section here, in which I give an argument to derive it from basic principles. I shall use the style that Einstein himself favored, that of a “thought experiment”, an idealized physical situation where it is easy to work out what must happen. I shall then back this up with a description of a real experiment, where the same basic feature is tested and verified. This set of experiments illustrates why physicists have such confidence in special relativity; indeed, special relativity is one of the best-tested theories in all of physics.

Many of the results are surprisingly simple to derive algebraically from Einstein’s basic principles. For readers who want to see how this works and how they relate to one another, Investigation 16.1 on page 201 contains the essential algebra and arguments.

Nothing can travel faster than light

Thought experiment

Imagine you are an experimenter using a linear particle accelerator, which employs strong electric fields to push a charged particle, like an electron, faster and faster in a straight line. In your experiment, you send a photon down the accelerator at the same time as you start accelerating an electron. Suppose that, a few moments later, the electron has reached the speed $0.999c$ with respect to you. (You are at rest on the ground.) From your point of view, the photon is now some distance ahead of the electron and still pulling away, but only gradually. Imagine now another experimenter who is flying past you with the same speed as that of the electron, and who measures the speed of the electron and of the photon. The electron’s speed is momentarily zero with respect to the flying experimenter, of course. And, by the invariance of the speed of light, this experimenter measures the photon’s speed to be c . So the electron, despite its enormous acceleration, has from this point of view not come one bit closer catching up with the photon! If you wait another few moments, until the electron reaches the speed of $0.999999c$, there will still be

In this chapter: we examine the foundations of special relativity in detail, deriving all the unusual effects from the fundamental postulates, examining the experimental evidence in favor of each one, and showing that the theory is self-consistent even if at first sight it seems not to be.

►The image under the text on this page illustrates length contraction. The top figure is after Leonardo da Vinci’s famous drawing. The bottom figure has the dimensions that an experimenter would measure if the experimenter were flying across the original drawing at a speed of $0.9c$.

In this section: how to understand that the invariance of the speed of light prevents anything going faster than it, or indeed even catching up with a photon.

▷ Interestingly, this argument on not accelerating up to the speed of light does not exclude from Nature the possibility that there are particles that simply start out at speeds faster than light. These logical possibilities are called **tachyons**. There is no evidence that they exist, and because they create problems with causality (see the section on the loss of simultaneity below) most physicists do not expect them to exist.

▷ The *synchrotron* gets its name from having to synchronize its forces with the exact position of the particle as it goes around faster and faster.

In this section: the invariance of the speed of light also means that photons can never come to rest.

another (imagined) experimenter who would measure this electron to be at rest and the photon still to be moving at the full speed c . If you think about this, you will see that, no matter how much you accelerate it, the electron won't reach the speed of the photon, because there is always a perfectly good experimenter for whom the two speeds are not even close! The only possible conclusion is that the electron simply cannot travel at the speed of light. And since it can't get to the speed of light, it can't go faster than light.

Real experiment

Linear accelerators like this one, and circular accelerators (called synchrotrons) that push electrons (or protons) around a circle, operate successfully every day. While they do not directly measure the speed of the electron relative to the hypothetical photon, they do something just as good. They need to anticipate exactly where the electron is at any moment so they can give it just the right push to keep accelerating it. If the electron did not turn up in the right place, as predicted by special relativity, then these enormous machines would simply not work. If an electron ever moved faster than light in such a machine, the experimenters would soon know it!

Light cannot be made to stand still

Thought experiment

Since every experimenter must measure light to have speed c , there is no experimenter for whom light can stand still.

Real experiment

Strictly speaking, what we are saying is that, if a photon or any other particle travels exactly at speed c , then it cannot go any slower, and so it cannot be made to stand still. This is established indirectly by the successful operation of particle accelerators, as just described. It is a separate question to ask if photons are such particles: do real photons travel with speed c ? This is equivalent to asking if they have a non-zero rest-mass. This is an experimental question about the nature of photons, and so far there is no evidence for any rest-mass. It is important to understand that, if experiments did show that photons had a small but non-zero rest-mass, then this would not upset special relativity. It would mean that we would no longer want to call c the speed of *light*, so we would call it something else, like the Einstein speed. But it would still be fundamental even if light happened not to follow it, and it would still be a barrier to the speed of all objects, including sluggish photons.

It is also important to understand that Einstein's principle applies to the speed that light travels only if the light is free to move without disturbance. Obviously, if a photon is reflected backwards by a mirror, then at the moment of reflection one might be tempted to say that it is "standing still". But in such a case we are not dealing with a photon on its own. It would not reflect if it were not interacting with the electrons and protons of the mirror. What actually happens on a microscopic level is that the incoming photon is absorbed by the electrons of the mirror, which are set into oscillation by the photon's oscillating electric field. The result is, for some materials (shiny ones), that the electrons' oscillation creates a new photon that moves away from the mirror in the opposite direction. The incoming and outgoing photons are free and move at speed c , but they are not the same photon, because at the "moment" of reflection (which actually lasts no more than a few oscillation periods, perhaps 10^{-14} s) there is no independent photon at all.

Not only mirrors, but also transparent materials undergo complex interactions with light. When "light" moves through water or glass, what actually moves is a composite wave in the electric fields of photons and of the atoms of the material. The incoming photon causes atoms to oscillate, which then disturb other atoms further along, which oscillate, and so on through the material. This wave of disturbance

moves at speed less than c , a speed that is usually called the “speed of light” in the material. But it is not the speed of a free photon, which is always c (provided photons are massless). If the frequency of the photon is high enough, the electrons of the material won’t be able to respond to it, and the photon will be able to travel freely through it. This is why X-rays penetrate most materials, and why the speed of light in any material always gets closer to c as the frequency of the photon gets higher.

Clocks run slower when they move

Thought experiment

We shall show this by considering a very simple kind of clock, one that just reflects a photon between two mirrors and “ticks” once for every round-trip that a photon makes. This is a good clock for us to use in a thought experiment, even if it is rather impractical to make, because it directly involves light, whose simple properties we completely understand. Suppose that the time it takes for light to go up and back when the mirrors are at rest is τ , the time for one tick of the clock. Now imagine the clock is moving with a certain speed v in a direction perpendicular to the line joining the mirrors. Now the light has to travel further on its round-trip. As Figure 16.1 on the next page shows, the light travels on the two sides of a triangle, whose base is the distance the clock travels during the round-trip travel time. Since the side of the triangle is longer than the distance between the mirrors, the total distance traveled by light in the moving clock is larger than in the clock at rest. By Einstein’s principle of the invariance of the speed of light, the photon travels at the same speed in each case, so it must take longer to go up and back in the moving clock.

The calculation in Investigation 16.1 on page 201 shows that the time it takes for the clock to tick if it moves at speed v is longer by the factor $\gamma = (1 - v^2/c^2)^{1/2}$ than the time it takes if it is at rest:

$$\Delta t_{\text{moving clock}} = \gamma \Delta t_{\text{clock at rest}}. \quad (16.1)$$

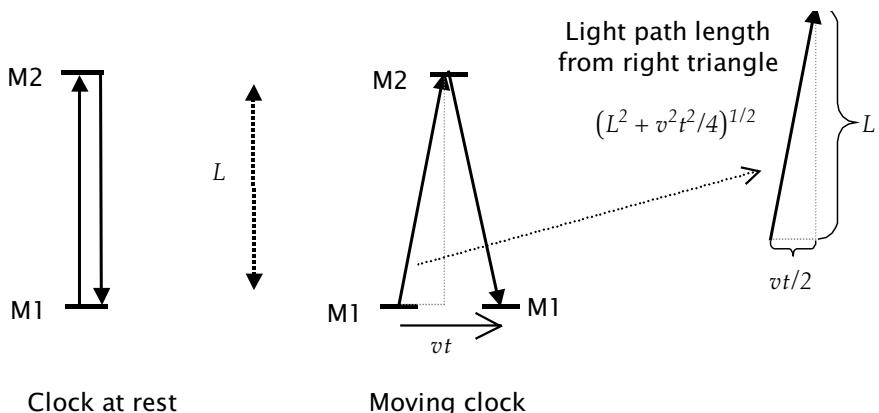
This is called the *time dilation* effect: clocks that move run more slowly, so time is stretched out (dilated). Note that this must happen to all clocks, not just ones based on light. If we have built a sufficiently accurate mechanical clock that keeps time with our light clock, and we place the two side-by-side and at rest, then they will remain synchronized for as long as we wish. We can arrange, for example, for them to give off flashes of light once every second, simultaneously. If another experimenter travels past us on a train, then our light clock has a non-zero speed with respect to the experimenter, and so it will run slowly as measured by the experimenter. This must then also be true of the mechanical clock, since we know it will emit a flash of light every time the light clock does; because these flashes happen right next to each other, the moving experimenter must see the two clocks flash at the same time, just as we do. The mechanical clock, which could be any other clock, therefore runs just as slowly as the light clock. A particularly important clock is our biological clock: a moving experimenter ages more slowly than one at rest!

Real experiment

Time dilation is easy to measure directly. Nature provides us with a number of natural clocks in the form of unstable elementary particles and nuclei. Particles decay at random: some decay rapidly and some take a long time. And they have no memory: a particle that has by chance lived 100 years is just as likely to decay in the next second as a particle of the same type that was created one millisecond ago. For this reason, the decay is fully characterized by one number, called the *half-life* of the particle type, a concept that we introduced in Chapter 11. It follows that, in a sample of N identical particles, there will be on average $N/2$ decays in the time τ .

In this section: the only way that light can have the same speed to all observers is if observers disagree on time and space measurements. Here we see that moving clocks must run slowly.

Figure 16.1. A simple clock based on reflecting light. This clock ticks once when a photon makes a round-trip from mirror M1 to mirror M2. When the clock is at rest (shown at the left in the figure) the photon travels a distance $2L$ for this. When the clock is moving (shown at the center), light must travel a longer distance in order to return to the mirror M1, which has moved since the photon left it. The geometry used to calculate this distance is shown at upper right.



So if we measure the number of decays per unit time in a sample of particles, we have a natural clock, a natural way to measure τ . Every time the number of atoms reduces by one-half, this natural clock ticks a time τ .

An observation of time dilation using such a clock was first made for particles called **muons**, which are produced abundantly in the upper atmosphere when high-energy cosmic rays strike oxygen or nitrogen nuclei. The half-life τ of a collection of muons at rest is 2.2×10^{-6} s. Even if the muons are produced moving at close to the speed of light, they could travel no more than 660 m in 2.2×10^{-6} s before losing substantial numbers to decay. Yet muons are detected easily at ground level, many tens of kilometers below where they are produced. And at the top of a 3000 m mountain, experiments see only a few more than at ground level, rather than the factor of $2^{(3000\text{ m}/660\text{ m})} \approx 23$ more that would be expected if half of them decayed every 660 m. Time dilation explains this: since they travel at nearly the speed of light, their internal clocks slow down dramatically, and they live much longer according to our clocks. Similar experiments can be done with unstable particles produced at high speeds in accelerators, and the predictions of special relativity are confirmed to a high accuracy.

A more practical application of time dilation today involves the Global Positioning System (GPS) that we discussed in Chapter 2 as an illustration of the gravitational redshift. This redshift produces considerable differences between the rates at which clocks on the ground and in orbit run, and these differences have to be corrected often in order for the navigation system to work. What we did not explain in Chapter 2 is that time dilation produces differences of a similar size, so that it too has to be calculated and removed with the gravitational redshift: an annoying but unavoidable nuisance for the navigation system! If special relativity were not right, we would quickly learn about it from the GPS.

Novices to special relativity often worry that the time dilation effect is inherently self-contradictory, and that this should show up in experiments. The worry goes as follows: if experimenter A measures experimenter B's clocks to run slowly, simply because B has a speed v relative to A, then the principle of relativity implies that B will also measure A's clocks to run slowly, since the speed of A relative to B is also v . But this seems to be a contradiction: how could B be slower than A and A be slower than B? This is an important question, and one that goes to the heart of understanding special relativity. I shall give considerable attention to it in a separate section on the so-called twin paradox, at the end of the chapter. But there is a brief answer that we can look at here and see that the appearance of a contradiction comes

from comes from comparing what are in fact two different measurements.

Let us look carefully at how each experimenter performs the measurement. For example, when A measures the rate of ticking of one of B's clocks, A must effectively use *two* of his own clocks: one to record the time where B's clock first ticked, and the second to record the time where B's clock next ticks. A needs two clocks because B's clock is moving. Thus, A's measurement involves a comparison of three clocks in total: two of A's clocks, which must run at the same rate (must be synchronized), and one of B's clocks. By the same reasoning, the experiment performed by B involves two of B's clocks and only one of A's. So although both experimenters describe their experiments as a comparison of one set of clocks with another, they actually compare different sets of clocks, so they are not doing the same experiment, and they do not need to get "consistent" results.

The length of an object contracts along its motion

Thought experiment

How can the muon result be explained to an experimenter traveling with the muons? Since they are at rest with respect to this experimenter, then half of them decay after only $2.2\ \mu\text{s}$. The Earth has been approaching the experimenter at nearly the speed of light during this time, but even so it cannot have traveled more than 660 m. Yet most of the muons have reached the ground. The inescapable conclusion is that the ground is less than 660 m from the top of the atmosphere, as measured by the experimenter moving with the muons. A length that is more than 20 km as measured by an experimenter at rest on the Earth has contracted to less than 660 m when measured by an experimenter moving at nearly the speed of light.

This effect is called the *Lorentz–Fitzgerald contraction*, because Lorentz and Fitzgerald were the first to propose that it and time dilation actually occurred. The formula is, following the pattern of earlier ones,

$$L_{\text{moving object}} = \sqrt{1 - v^2/c^2} L_{\text{object at rest}}. \quad (16.2)$$

But there is a crucial difference between what Lorentz and Fitzgerald predicted and what Einstein showed really happens. For Lorentz and Fitzgerald, the speed v in this formula was the speed of the object through the ether. Thus, in the Michelson–Morley experiment, they expected that the length of the arm of the interferometer that lay along the direction of motion of the Earth was physically shorter than the other arm, but that this was unfortunately unmeasurable: as soon as Michelson held a ruler up against this arm to measure its length, the rule would contract by the same amount, so the arm would appear to have its rest-length. Nevertheless, to Lorentz and Fitzgerald, the arm "really" was shorter. For Einstein, the length *is* what the experimenter measures, and the contraction occurs only when the object moves relative to the experimenter who makes the length measurement. This fundamental difference in interpretation is the main reason that Einstein gets the credit for discovering the contraction effect, even though the mathematical expression is the same as for Lorentz and Fitzgerald, and even though we honor their contribution by naming the effect after them.

Real experiment

We have already noted above that unstable muons must see the Earth's atmosphere greatly contracted, in order for them to reach the ground in their decay lifetime. The same effect occurs for elementary particles in accelerators; from their point of view, the accelerator tube must be very short, so that they don't decay in the middle of it. It must not be thought that the Lorentz–Fitzgerald contraction is somehow an illusion produced by a problem measuring time or by clocks that don't behave

In this section: just as time slows down with speed, so also lengths contract along the direction of motion.

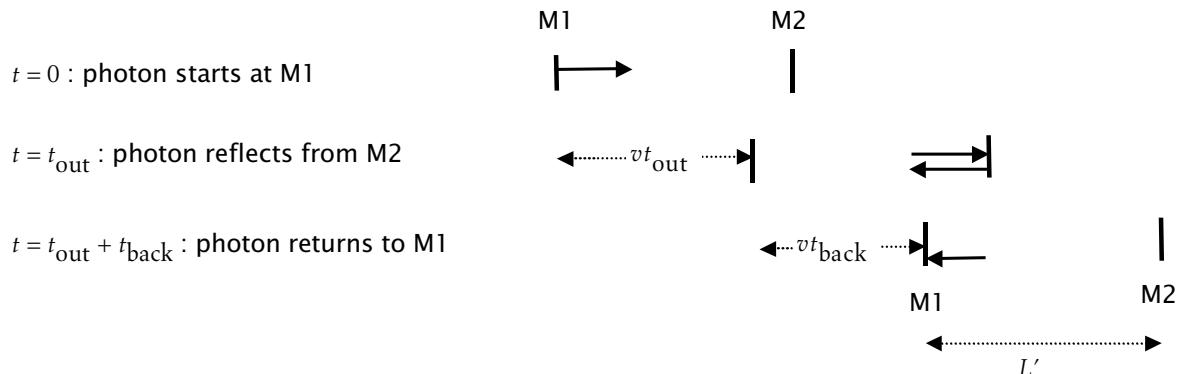


Figure 16.2. The simple clock of Figure 16.1 on page 198, now moving along its length. Three snapshots are shown, at the moments when the photon starts out from mirror M1, reflects from M2, and returns to M1. The distance between the mirrors is called L' , to allow it to be different from the distance L when they are at rest. On its way out, the photon travels much further than L' , because the end mirror is moving away from it. On the way back, it travels less than L' . Its total round-trip time must be the same as for the moving clock in the previous figure, since the two are at rest with respect to one another and so must stay in synchronization.

well: it is a real contraction. For example, if one tries to accelerate a solid body, then its contraction requires an extra input of energy to squeeze the atoms of the body closer together. This extra energy goes into the total mass of the moving body. Without it, the mass would not increase in proportion to γ , as we saw above that it must.

Loss of simultaneity

In this section: the disagreements over time measurements lead to a breakdown of the idea that there is a universal time. Any single experimenter can measure that a given pair of events occur at the same time; but to another experimenter, one event will occur before the other. Neither experimenter is "wrong".

According to relativity, we simply have to give up the idea that it is possible to say that two things happened at the same time and expect everyone to agree. If they happen at the same time *and* at the same place, then everyone will agree about this. (We used this for our two flashing clocks above.) But if the events are separated, the notion of simultaneity is not universal.

Thought experiment

A simple thought experiment involving two of our light-clocks will give us a direct example of how the notion of simultaneity depends on the experimenter. Let us put the two clocks next to each other on a table, at rest, but oriented so that their arms are perpendicular to one another. Let the ends of the clocks where the photons start out be in the same place. Now suppose the photons in each clock start out at the same time. The clocks are identical, except for orientation: the photons set out in perpendicular directions. Because the arms have the same length, the photons reach the end mirrors at exactly the same time. These reflections are *simultaneous* to us. But of course they occur in different places.

Now we shall see that they cannot be simultaneous to an experimenter who watches the same thing happen as the clocks move. Suppose our laboratory is in a train, and this is traveling at speed v along the direction of one of the clocks. Then we know that the experimenter on the ground will measure our clocks to be running slowly, but still both go at the same rate: the total time it takes a photon to go out and back in both clocks is the same. Here, however, we are interested only in the first part of the photon's journey, going out. For the clock perpendicular to the motion of the train, the journeys out and back are identical, so the reflection event occurs at exactly one-half of the time of a "tick". (This is the situation shown in Figure 16.1 on page 198.) For the clock oriented along the direction of motion, the outward and return journeys of the photon are not identical. On the outward leg, the distant

Investigation 16.1. The light-clock shows how time and space warp at speed.

The time dilation follows directly from the principle that the speed of light is the same to all observers. We will use the light-clock illustrated in Figure 16.1 on page 198.

First, let us agree that the clock at rest, on the left in the figure, is a good clock, and once calibrated it will run at the same rate as any other clock we could construct, held at rest with respect to us. So if, as we shall show, the clock runs at a different rate when moving, then this will apply to all other clocks, including the psychological perception of time.

The clock ticks once each time the photon returns to the bottom mirror M1. While the clock is at rest, this takes a time $\tau = 2L/c$. Now let the clock move, or equivalently let a moving experimenter measure the time it takes for the same clock to tick, using an identical light-clock at rest in his own laboratory. All we need to compute is the distance light travels in one tick. From the figure, it is clear that it travels along a triangular path because the mirror M1 is now moving and it has to meet it after being reflected at M2. If the return trip takes time t , the figure shows that the path has total length $2(L^2 + v^2 t^2/4)^{1/2}$.

Now, the photon traveled at speed c , so the time t it took is this distance divided by c . This gives us an equation to solve for t :

$$t = \sqrt{L^2 + v^2 t^2/4}/c.$$

Squaring this gives

$$t^2 = 4L^2/c^2 + v^2 t^2/c^2,$$

which can be solved for t . We shall call this value of t the ticking time of the moving clock, τ' :

$$\tau' = \frac{2L}{c} \sqrt{1 - v^2/c^2} = \gamma \tau,$$

where γ was defined in Equation 15.4 on page 188. Since γ is larger than 1, the tick time τ' of the moving clock is larger than that of the clock at rest, so the moving clock is running slowly.

From this we can also derive the Lorentz-Fitzgerald contraction: it is the other face of time dilation. Consider the muons created in the upper atmosphere and heading for the Earth, as described in the text. Because of time dilation, they live much longer than their half-life as measured when they are at rest ($2.2\ \mu\text{s}$), allowing them to travel much further, so that many reach the ground.

Now, how can this be explained to an experimenter traveling with the muons? Since they are at rest with respect to this experimenter, then half of them decay after only $2.2\ \mu\text{s}$. The Earth has been approaching the experimenter at nearly the speed of light during this time, but even so it cannot have traveled more than $660\ \text{m}$ in this time. Yet the muons have reached the ground. It follows that the ground must be less than $660\ \text{m}$ from the top of the atmosphere,

Exercise 16.1.1: Practical time dilation

An airline pilot spends $20\ \text{h}$ per week flying at a speed of $800\ \text{km h}^{-1}$. Over a career of 30 years, how much younger is he than if he had never flown?

Exercise 16.1.2: Exploring the Galaxy

A future civilization manages to construct a rocket that will move through the Galaxy with $\gamma = 10^4$. What speed does it go? How long will it take to cross the Galaxy, some $20\ \text{kpc}$? How wide is the Galaxy as measured by the voyagers?

Exercise 16.1.3: The dangers of space exploration

Show from Equation 15.7 on page 192 that a photon with an energy E as measured by an experimenter at rest in the Galaxy, and which is approaching the spaceship of the previous exercise head-on, has an energy relative to the spaceship of $E' = 2\gamma E$. Suppose the spaceship is approaching a normal star like our Sun. Use Wien's law (Equation 10.9 on page 117) to infer that the radiation reaching the spacecraft from the star will have an effective temperature of 50 million K! That means it would be a source of gamma-rays, and the inhabitants of the spacecraft would have to shield themselves carefully.

as measured by the experimenter moving with the muons. A length that is more than $20\ \text{km}$ to an observer at rest on the Earth has contracted to less than $660\ \text{m}$ by moving at nearly the speed of light.

We can deduce what the size of this contraction should be by simply turning the clock we used for the time dilation thought experiment on its side, so it is now moving at speed v along its length. This cannot change its ticking rate, since the clock is at rest with respect to the moving clock we computed above, just turned in a different direction. An observer at rest with respect to them would certainly expect them to maintain synchronization. From this we will see that its length must change. We can see the situation in Figure 16.2.

On its way from mirror M1 to M2, the photon takes a time t_{out} . In this time it has to travel the length of the clock, L' (not necessarily equal to L), plus the extra distance that M2 has moved in the time t_{out} , which is vt_{out} . It travels at speed c along this path (Einstein again), and so we again have two expressions for the distance it traveled, which must be equal:

$$ct_{\text{out}} = L' + vt_{\text{out}},$$

from which we can deduce that

$$t_{\text{out}} = L'/(c - v). \quad (16.3)$$

Similarly, the return journey is shorter because the mirror M1 is now catching up with the photon. So a similar expression for the distance traveled in terms of the time t_{back} for this journey is

$$ct_{\text{back}} = L' - vt_{\text{back}} \Rightarrow t_{\text{back}} = L'/(c + v).$$

Now, these two times must add up to the round-trip time we had when the clock was oriented perpendicular to its motion, which was $2L/c(1 - v^2/c^2)^{1/2}$. This determines L' in terms of L :

$$\begin{aligned} \frac{L'}{c - v} + \frac{L'}{c + v} &= \frac{2L}{c} \frac{1}{\sqrt{1 - v^2/c^2}} \\ \Rightarrow L' &= L\sqrt{1 - v^2/c^2}. \end{aligned} \quad (16.4)$$

We see that the length of the clock when it moves along its length is shorter than its resting length, by the factor $(1 - v^2/c^2)^{1/2}$. This is the only way to keep the clock ticking at the same rate, independent of orientation, and to have the speed of light the same for all experimenters, regardless of their state of motion relative to one another. It is important to keep this last fact in mind: if Newton had been doing this calculation he would have obtained a very different result, because for him the speed of light would have been different in different circumstances; where we always simply used c , he would have used different and more complicated expressions. But we know from experiment that Einstein was right and Newton wrong on this.

mirror is running away from the photon, while on the return journey the mirror at the back of the clock is approaching the photon. The outward journey must, therefore, take longer than the return journey, which means it must take longer than half a “tick” of the clocks. So the reflection event in this clock will be later than that in the other clock. *Two events that are simultaneous to one experimenter are not necessarily simultaneous to another.*

Simultaneity is therefore not a universal property of pairs of events. Two events that are separated in space may be simultaneous to one experimenter and not to another. In fact, we see in this example that the event that is more to the rear of the train, which was the reflection along the line perpendicular to the motion of the clocks, is the one that occurred first. This is a general property: if two events are simultaneous to one experimenter, and that experiment is viewed by another experimenter who is moving, then the event toward the rear will happen first. Notice that events separated perpendicular to the direction the train is moving remain simultaneous: this effect applies only in the direction of motion.

This is an inescapable consequence of the way the speed of light behaves: if the speed of light were to combine with other speeds in the way that Galileo and Newton expected, then the experimenter on the ground would agree with the one on the train that the two events were simultaneous. Because light does not change speed when viewed by the different experimenters, simultaneity depends on who measures it.

But the same property of light implies that some pairs of events will have an invariant time-ordering. These are events that can be connected by a photon or, indeed, by something traveling slower than light. Here every experimenter can observe the event where the photon started and the event where it finished, and clearly (since light always travels at speed c) the time between the two events will be non-zero. Pairs of events that can be connected by a single photon are said to be lightlike-separated. If the events occur even closer to one another, so that a particle traveling at less than c can go from one to the other, we say they are timelike-separated. Events that occur too far apart to be connected by a photon are said to be spacelike-separated.

All the effects of special relativity that we have studied – time dilation, Lorentz–Fitzgerald contraction, the loss of simultaneity – are related to each other. We saw above how to relate the Lorentz–Fitzgerald contraction to the time dilation effect. Let us here derive the Lorentz–Fitzgerald contraction from the loss of simultaneity. The first question to answer, and in one sense the deepest, is: how can we, at least in principle, measure the length of something that moves, say the train of the previous example? This requires us to compare its length with the length of a ruler or other length standard that is at rest with respect to us. How do we compare a moving length with one at rest?

One acceptable way is to imagine the train moving past a wall. If we mark the locations of the front and rear of the train on the wall, we can then measure the distance between the marks at leisure to get the length of the train after it has passed. Of course, if we adopt this method, we must insure that we mark the front and rear *at the same time*. It would not make sense to mark the location of the rear at one time and then wait a minute to mark the location of the front: the train would have moved in the meantime, and the marks would not be separated by the true length of the train. They have to be marked at the same time.

We see, therefore, that measuring the length of a moving train in this way requires us to use the notion of *simultaneity*. Since simultaneity depends on the experimenter, so does the length of the train.

It will reward us to look a bit more closely at how this works. We have just seen that a clock at the back end of the train runs ahead of the one at the front, as far as we (experimenters on the ground) are concerned. Therefore, we do not make the marks when these two clocks read the same time, say 10:00. Instead, if we happen to mark the front of the train when its clock reads 10:00, then we must mark the rear when its clock reads a bit *later* than 10:00. From the point of view of the experimenter on the train, we have not done it right: we have waited too long to mark the rear, and therefore we have obtained too short a length. The rear has caught up a little with the front during the extra time we allowed at the back. But from our point of view, we have done the right thing: we have marked the two ends at the same time, as we measure time. And this leads, of course, to a length that is smaller than the length measured by the experimenter on the train: the Lorentz–Fitzgerald contraction.

The relativity of simultaneity therefore leads directly to the relativity of lengths. Our argument shows that it happens only to lengths oriented along the direction of motion: because simultaneity holds in perpendicular directions, perpendicular lengths do not contract.

The notions of timelike and spacelike separations, which we introduced above, are related to some simple ideas about where and when events occur. If two events are timelike-separated, then by definition there is a particle moving at less than the speed of light that goes from one event to the other. Suppose we jump on a spaceship traveling at this speed. If this particle is at one place next to the spaceship when the first event occurs, then it will be at the same place (in relation to the spaceship) when the second event occurs, because the particle is not moving relative to the spaceship. It follows that, if two events have a timelike separation, there is a class of experimenters who will measure the positions of the two events in space to be the same: they occur in the same place at different times for these experimenters.

By analogy, it is not hard to see that, for any pair of *spacelike*-separated events, there is a class of experimenters who will see them as simultaneous: they occur at the same time in different places.

This has implications for the particle called the tachyon, which we mentioned earlier. If it travels faster than the speed of light, then it can travel from one event A to another B that is spacelike-separated from the first. There will be some experimenters who will measure these two events to occur at the same time, and for these experimenters the tachyon travels infinitely fast. For some other experimenters, the event A at which the tachyon started occurs *after* the event B at which the tachyon finished (the ordering of spacelike-separated events in time depends on the experimenter's state of motion), so for these experimenters the tachyon actually travels backwards in time. For this reason, it is hard for most physicists to believe that tachyons exist: they would seem to upset all our notions of cause and effect.

Real experiment

Simultaneity is important for the GPS that we have already mentioned several times. In order to provide a receiver on the ground with the correct information from which to deduce its position, the clocks on the satellites to be coordinated in some way. Suppose we were setting up this system and tried to arrange for the clocks in the satellites to be synchronized with each other. Consider a pair of satellites, one following the other in orbit around the Earth. If they synchronize with each other, then from the point of view of an observer on the ground, the clock in the leading satellite will be behind that in the following one. This is not acceptable: they should be synchronized as measured by ground-based experimenters. This means that, when the satellites pass time signals among themselves, they must be aware of the fact that their own clocks are not synchronized with one another. If the clocks

►The situation for the GPS is slightly more complicated, since the satellites orbit in a gravitational field, and this means that gravitational effects on clocks, like the gravitational redshift, must also be taken into account. But the principle of the discussion here is not changed by these complications.

In this section: the only way that objects can fail to be accelerated past the speed of light by applied forces is if their mass increases so rapidly that the force produces less and less acceleration.

did not behave as relativity requires, then we would soon measure this. The GPS system is, therefore, a continuous demonstration of the relativity of simultaneity.

The mass of an object increases with its speed

As we described earlier when we introduced this idea, by “mass” we mean the usual symbol m that appears in Newton’s second law, $F = ma$. This is called the inertial mass of the object. But because the mass changes with speed, this form of Newton’s law is not adequate. We saw in Chapter 6 that when the mass of an object changes as it moves, we need to replace the simple version, $F = ma$, with the equation that we called the rocket equation: the product of the applied force with the time it acts equals the change in the momentum of the particle. Since the speed of the particle hardly changes when it is already near c , the increase in the momentum must be almost entirely due to an increase in the mass of the particle.

Thought experiment

It is not hard to derive the formula for the mass of a moving particle from the formula for the relativistic Doppler shift, Equation 15.7 on page 192 above, and the equivalence of energy and mass, $E = mc^2$. The derivation is instructive because it shows that the increase of mass with speed is a direct consequence of time dilation. Since mass and energy are proportional, this is another illustration of the deep connection between time and energy that we first met in Chapter 6.

Imagine a simple physical event, where a particle of rest-mass m_0 suddenly decays into two photons of equal energy. (A particle called the π^0 does this.) The particle disappears altogether, and the photons travel away from it in opposite directions. Seen by an experimenter who was at rest with respect to the particle before it decayed, the two photons have equal energy, $m_0c^2/2$. That is all we need to know about this event.

Now suppose an experimenter watches the same event while speeding past it at speed v , and suppose that this speed is in the direction that one of the photons takes after the decay. Then the initial energy of the particle depends on its total mass m_1 . Let us forget that we have a formula for m_1 , and try to derive it by conservation of energy, using the energy of the two photons as measured by this experimenter. The two photons now come off with different energies. The one that is going in the same direction as the experimenter is redshifted. Its frequency as measured by the experimenter at rest was its energy divided by Planck’s constant h , $f_0 = m_0c^2/(2h)$. Its frequency as measured by the moving experimenter is, by the relativistic redshift formula Equation 15.7 on page 192, $f_0\gamma(1 - v/c)$. Its energy is h times this.

The other photon is going in the other direction, so it is blueshifted. Since it has the same frequency f_0 with respect to the first experimenter, its frequency with respect to the moving experimenter is $f_0\gamma(1 + v/c)$. Again, its energy is h times this. The total energy, therefore, as measured by the moving experimenter, is $2hf_0\gamma$: the Doppler factors of v/c have just cancelled. Putting back f_0 into this gives us a total energy of γm_0c^2 . By energy conservation, this must be the total mass-energy of the particle before it decayed, or m_1c^2 . We find from this the formula quoted earlier without proof, that the inertial mass is $m_1 = m_0(1 - v^2/c^2)^{-1/2}$.

Where has the factor of γ come from? It is the new factor of γ in the relativistic Doppler formula. This arose, as we saw above, from time dilation. It is the same factor that causes the transverse Doppler effect. This shows again the deep relationship between time and energy that we first mentioned at the end of Chapter 6.

Real experiment

Verification of the increase of inertial mass with speed comes again from the synchrotron accelerator. Keeping the accelerated electron on the circular track requires the machine to produce precisely the right acceleration, even as it goes faster and

faster. This is done by calculating the force needed to produce that acceleration in a particle whose mass depends on speed in just the way Einstein's theory predicts. If this prediction were wrong, accelerators would simply not work.

The record for a particle moving close to the speed of light is not held by a manmade particle in an accelerator, but rather by a cosmic ray. Protons regularly hit the upper atmosphere at high speeds, and are called cosmic rays. They produce the showers of muons that we used to illustrate time dilation. By measuring the muons and other particles produced by the collisions of cosmic rays with oxygen and other nuclei in the atmosphere, astronomers can infer the speed and mass of the incoming proton. The largest energy so far measured is about 10^{21} eV, or about 160J. This one elementary particle carried as much energy as your body would extract from eating two spoonfuls of sugar! The mass equivalent to this energy (see the next section) is about 10^{12} times the rest-mass of the proton. That means that the proton was traveling at a speed of $0.99999999999999999999995 c$ when it hit the Earth!

Energy is equivalent to mass

Thought experiment

Remember that energy is conserved, but it can be converted between different forms. It follows that *any* form of energy put into a particle contributes to its inertial mass. For example, if I begin with two protons, and squeeze them together against their mutual electric repulsion, then the force I have exerted to push them together has done work and put energy into the system. If I then hold them together somehow, I have a system with more energy than I started with. Its mass must be larger than the masses of the two protons alone. If the system is just sitting at rest somewhere, then this mass is its rest-mass. By allowing the two protons to fly apart, I convert some of this rest-mass back into energy, the kinetic energies of the particles. Therefore, even rest-mass, at least for composite particles, is convertible into other forms of energy. Moreover, this is not a mysterious process: one is just releasing energy that was put in when the composite object was assembled. The law of conservation of energy must include rest-mass energy. This is $m_0 c^2$ for a particle of rest-mass m_0 .

Notice that, apart from the factor of c^2 , mass *is* energy. Rest mass is one part of the energy of a system, but its total mass is its total energy divided by c^2 . All forms of energy in a composite system contribute to its rest-mass, and a moving system has a mass that is greater than its rest-mass. This extra energy can be called its kinetic energy, but it is *not* given (except for slowly moving particles) by the Newtonian formula $mv^2/2$. We will find the correct formula below.

Real experiment

When the electron accelerated in a real synchrotron smashes into a target, which is what high-energy-physics experiments usually require, very sensitive and fast measuring machines measure the energy released, in terms of the rest-masses and kinetic energies of all the particles produced in the target. This always totals the energy put into the electron's total mass by the forces that accelerated it.

This law holds to such accuracy that it can allow physicists to discover new particles. The neutrino, which we learned about in Chapter 11, was first noticed this way: the energy (and momentum) in the particles that were identified as having come from the decay (splitting up) of a particular initial particle did not add up to the energy and momentum of the initial particle, even once allowance had been made for the energy in its rest-mass. In 1934, following an earlier suggestion by Pauli, Enrico Fermi showed that the deficits in decays of similar particles could always be made up by postulating that there was an undetected particle that traveled

In this section: the close relationship between energy and time leads in relativity to a link between energy and time dilation, so that the faster an object goes, the more energy it has. And this energy is exactly proportional to its inertial mass.

▷We met both Pauli and Fermi in Chapter 12.

at the speed of light and had no electric charge. He named the particle the “neutrino”, which means, in Italian, a small neutral particle. As we have seen, it took roughly twenty years to develop the technology to make detectors sensitive enough to register neutrinos directly.

Nuclear reactors and nuclear weapons are, of course, the standard examples of the conversion of rest-mass into energy. In these devices, a composite particle – often a nucleus of uranium or plutonium – is split into two smaller particles whose rest-masses total less than that of the first. The excess rest-mass appears as the kinetic energy of the two smaller products, and this is the source of energy for the device. The hydrogen – or “thermonuclear” – bomb works the other way, by fusing hydrogen nuclei to form helium, which also has a lower rest-mass than that of the “raw material” nuclei. As we saw in Chapter 11, this is also the power source in stars.

Photons have zero rest-mass

In this section: an object with non-zero rest-mass would have an infinite energy if it could be accelerated to the speed of light.

Since light itself carries a finite energy, the rest-mass of a photon must vanish. This also leads to the fact that the momentum of a photon is proportional to its energy.

Thought experiment

The fact that a photon must have momentum as well as energy follows from the same thought experiment that we used above to derive the dependence of the mass on the particle’s speed. As measured by the experimenter at rest with respect to the particle that decays, there is zero total momentum because the particle was not moving. But now look at the decay of the same particle from the point of view of the moving experimenter. The particle has initial speed v , so initial momentum $m_1v = \gamma m_0 v$. After the decay, where does this momentum go? If momentum is conserved, which is a fundamental principle we don’t want to give up, then the photons must carry away the momentum as well as the energy of the particle. Einstein showed that the momentum p of a photon is related to its energy E by

$$p = E/c. \quad (16.5)$$

► We shall use the usual physicists’ symbol p for momentum here. Don’t confuse this with pressure.

The context of the discussion should always make it clear which quantity we mean.

It is easy to show that this formula is exactly what is needed in this case to give momentum conservation. The momentum of the forward-going (redshifted) photon is $(hf_0/c)\gamma(1 - v/c)$, and the blueshifted one has momentum $-(hf_0/c)\gamma(1 + v/c)$, which is negative because the photon is going backwards relative to this experimenter. Added together, these give a total momentum of $-2hf_0\gamma v/c^2$. Putting in $f_0 = m_0c^2/2h$, we find that the total momentum is $-\gamma v m_0$. Since, before the decay, the particle was moving backwards with speed v with respect to this experimenter, and since its mass was γm_0 , this total momentum is exactly the momentum the particle had before the decay. Einstein’s formula, Equation 16.5, is just what is needed to insure conservation of momentum.

In fact, this formula is not very mysterious. It is a special case of the general expression for momentum, $p = mv$. This relation holds in relativity just as in Newtonian mechanics provided we use the total inertial mass of the particle. Now, since the inertial mass is just the total energy E divided by c^2 , we also have that the momentum is $p = Ev/c^2$. Now, a photon has $v = c$, so for it we find $p = E/c$, as above.

Real experiment

Real accelerator experiments have to take into account both the energy and the momentum of any photons (gamma-rays) emitted in a reaction. Using the rules that the photon has no rest-mass and has a momentum proportional to its energy, such calculations always give consistent answers.

A more direct demonstration of the momentum of a photon is the *radiometer*, a device consisting of four vanes, each painted black on one side and white on the other, able to spin about its axis in a vacuum. When light strikes it, the black sides

absorb the light and its momentum, but the white sides reflect the light and therefore give the outgoing photons new momentum. This means that the force on the white sides exceeds that on the black, and the device spins appropriately.

Massive giant stars are also an example of the effects of the momentum carried by light: radiation pressure is their main support against gravity, and this is nothing more than the exchange of outward momentum from photons to gas particles, preventing the gas from falling inwards under gravity. And as a final example, recall our derivation of the Chandrasekhar mass in Chapter 12, which computed the momentum of relativistic electrons from the formula for the momentum of the photon. Gravitational collapse, supernovae, pulsars, and black holes all owe their existence at least partly to the fact that a photon's momentum is proportional to its energy!

Consistency of relativity: the twin paradox saves the world

In the section above on time dilation, I described why it is not a contradiction that both experimenters A and B measure each other's clocks to be going slowly. I pointed out that they actually perform different experiments. You might feel this is a little unsatisfying, because all I showed was that there does not have to be an inconsistency in relativity; I didn't really prove consistency.

There is in fact a much more subtle way to try to construct a situation in which special relativity looks self-contradictory. This is usually called the "twin paradox". Because it is clever and it really brings out our conceptual difficulties with relativity, I will describe it here. But note from the start: it fails. It is not a true paradox at all. Relativity comes through it perfectly consistently.

The idea is to get away from an experiment that has to use three clocks to measure the rates of time of two different experimenters. Here is a version of the twin paradox that does not actually involve twins.

By the year 2202, overpopulation so threatened the Earth that the government of the (united) planet decided on a radical, but humane, solution. The exploration of the nearby part of the Galaxy for planets similar to the Earth had been fruitless: no place was known where excess Earthlings could be sent to live. Food was becoming a serious problem: almost all the land was used for dwellings, and the oceans had been over-farmed.

Their solution was ingenious: all of Earth's 100 billion people were distributed among 100 million spacecraft (constructed by solar-powered robots that mined the Moon), and each such community of 1000 people was assigned a list of target stars to visit in an attempt to find a new home. Their instructions were to stay at any star that turned out to be suitable, and to broadcast their discovery around the Galaxy so that other communities could go there. (In fact, many communities decided in secret that if they found a good new planet they would never tell anyone else!) If a star had no suitable planets, the community was to return to Earth immediately, re-stock their food supply (which had meanwhile been grown, stored, and packaged by robots), take a one-year vacation, and then head out for their next target. On average, it would take 1000 trips by each community before every star in the Galaxy had been visited.

It was only because of special relativity that this solution could work. At the speed of light, a spacecraft would take up to 60 000 years to cross the Galaxy and return. Yet the people on board would experience such a large time dilation that they would age very little. In practice, for the technology then available, the spacecraft produced an average time dilation factor of $\gamma = 10^4$, so they aged no more than six years on their round-trip. Communities that had nearer target stars returned after even less on-board time. And the plan was made attractive by the one-year vacation between trips: the ratio of this vacation to the length of a trip was better than the

►Do not confuse the radiometer with a device sold in toy shops that looks similar to it, but which spins in air. There the situation is complicated by the heating of the air, which is stronger near a black side, and which can cause the device to spin in exactly the opposite direction from that which would result from light pressure alone!

In this section: to test the logical consistency of special relativity, we confront the apparent contradiction in time dilation. If one experimenter sees another's clocks to be running slowly, how can the other experimenter measure that the first one's clocks are also going slower?

ratio of vacation to work-time that most of the population were entitled to in their Earth jobs.

The beauty of this strategy was that it solved the Earth's food shortages *even if no community ever found another Earth-like planet!* The reason was that, while the various communities were away, the Earth had many years to grow enough food to give the travelers when they returned. Since time for the spacecraft population was slowed to 10^{-4} of its rate on Earth, they ate very slowly while away (as measured by Earth clocks)! Time dilation was the perfect appetite-suppressant!

The plan was accepted by the people of the Earth, in most cases reluctantly, and they began to organize their small communities for the first set of trips. But then the political consensus was threatened in a dangerous way. A demagogue who was a fiery and persuasive speaker but a poor physicist began to build up opposition to the plan. Here is what the demagogue claimed.

Consider, instead, time dilation from the point of view of the people in the rockets leaving Earth. Once they reach their steady cruising speed, they are perfectly good experimenters, and when they compare the rate of time on Earth with the rate in their spacecraft, they will see that time on Earth is going slowly. From their point of view, Earth is receding from them at nearly the speed of light and hence suffers an enormous time dilation. Instead of the Earth being the place where food would grow for thousands of years before the community returned, it was really the other way around: the community would eat up its six-year on-board food supply and return to an Earth that had been growing food for only about five hours!

More sober politicians replied that it obviously could not turn out both ways: either the Earth had aged more than the communities when they returned, or they had aged more than the Earth, but not both at once. Then, since the Earth just sat around while the communities zipped around the Galaxy, it was clear that the communities were the ones that were the travelers and suffered time dilation, and not the Earth.

The demagogue replied that relativity claimed that all experimenters were equal, none was better than any other. It was a democratic theory, and there should accordingly be a vote to decide which point of view was right. In any case, how could anyone believe in time dilation at all when it gave two conflicting results? Maybe the right answer was that there was no dilation: the communities would have aged just as much as the Earth on their return, and since that could be as much as 60 000 years, nobody would live long enough to return. The one-year vacation was a bad joke, and the whole plan was a conspiracy by high government officials to get rid of everyone and then turn their own communities' spacecraft around to come back to a depopulated Earthly paradise. Many people began to believe that the demagogue was right.

Despite the polemic style, the demagogue had a point: time dilation is reciprocal, so how can we tell whether the Earth or the small communities will have aged more by the time they meet up again? Does this really indicate a logical flaw in the theory, as the demagogue claimed?

The key to answering the demagogue is to demonstrate that the two "experimenters" are not really on the same footing. The Earth-based experimenter is fine: clocks on the Earth can be constructed, synchronized, and run for the thousands of years required to see the communities return. But the communities are not ideal experimenters. In particular, they have to turn around. This changes their way of measuring time.

A good way to see the effect of this is to imagine a community that has a schism just before reaching their target star. Half of them do not want to tell anyone else

if they find a good planet, and the other half do. So only half of them (the honest ones) stop at the star, while the remainder (the secret ones) continue on.

Once the honest ones find that there is in fact no suitable planet at this particular star, they set off on their return journey. But now there is a huge time dilation between the two halves of the original community, since they are traveling at high speeds in opposite directions. Their relative speeds are now even larger than the speed of each relative to Earth. In fact, it is possible to show that the time dilation factor γ between the two groups is twice the *square* of the time dilation factor between each of them and the Earth, provided all speeds are close to the speed of light.

From the point of view of the secret group that continues without stopping (and which is therefore as good a set of experimenters as the people who never left Earth), time on the honest returners' clocks suddenly begins to go incredibly slowly, much more slowly even than time on Earth. From the secret group's point of view, Earth's clocks soon overtake the clocks of the honest returners, and by the time the returners get back, they find they have aged much less than the Earth. Of course, the returners may still not be expecting this, but by turning around they have changed their definition of time in such a way that they no longer agree with the secret group that they had agreed with before: the two groups begin at that point to have different ideas on simultaneity, for example, as well as on honesty.

So the returners are not good judges of what to expect about the behavior of time, and their expectations should be discounted. The demagogue was wrong to rely on their definition of time.

Rather than tell you right away how the political crisis on Earth turned out, I will give you the chance to decide your own ending to the story. Is the argument about the community that splits convincing enough to have defeated the demagogue? Correct physics does not necessarily win votes. My own ending to the story is upside down in Figure 16.3.

Because there is no contradiction between the time dilation measured by different experimenters, there is also no contradiction between consequences of time dilation, such as the Lorentz–Fitzgerald contraction. One can find books full of intellectual puzzles called “paradoxes” of special relativity, but in each case the challenge is to see how the wording of the puzzle leads one into thinking wrongly that the theory is self-contradictory. There are no real paradoxes in relativity.

Here is what happened between the demagogue and ordinary people were invited to the same place. Ordinary people were invited to come to the laboratory and measure the rate of radioactive decay. When they went into the accelerator, the initial radioactivity was high. When an abstract argument, such as the one given in the text, might not sway enough people, it therefore decided to demonstrate the appropriate form of time dilation experimentally. It constructed a circular particle accelerator and a source of muons. The source produced muons in bursts. Afterwards were fed into the accelerator, which took them to a gamma factor of about 10, or into a “muon cooler”, which reduced their speeds with respect to the demagogue had done the physics they had, at the last minute, saved the population of the Galaxy. Experimenters pride themselves for having demonstrated the correctness of the theory that the government guessed went down dramatically. The government guessed correctly: the demonstration convinced the population that the demagogue had done the physics wrong. The communities formed and began their life on a good planet of about 10¹⁰, or into a community of people going to other stars: they travelled at the same speed and kept returning to everybody. She died, of course, before anyone returned.

Figure 16.3. This is my ending for the story of how the Earth saved its population from starvation.

Relativity and the real world

In this section: many aspects of the world we live in depend on special relativity. We would simply not be here if stars could not convert rest-mass into energy, if stars at the ends of their lives could not explode.

Relativity is not a mere intellectual game, stimulating though it may be. We have seen from the experiments described in this chapter that it plays a central role, not only in physics experiments like big particle accelerators, but also in practical navigation systems and in nuclear power generators.

In fact, our lives depend on special relativity: if the Sun could not convert some of its rest-mass into energy, we would simply not exist. Our evolution has also been critically affected by special relativity at many times in our cosmological history. The protons, neutrons, and electrons of which we are made themselves came into existence about three minutes after the Big Bang; before that the Universe contained only a hot plasma of material in which particles were constantly being converted into photons and photons back into particles – rest-mass into energy and back again, in equilibrium. But, as we will see in Chapter 25, the Big Bang gave us only a Universe of hydrogen and helium, and not much in the way of heavier elements. The important elements of our lives – oxygen, carbon, iron, silicon, nitrogen, phosphorus, sodium, sulfur, chlorine, aluminum, lead, copper, zinc, silver, gold, uranium, and more – were all formed in stars as by-products of the conversion of matter into energy, and in fact we have seen that many of them were formed in the explosion of a big star (a supernova explosion), which happened before our Sun formed and which mixed the new elements into the gas cloud from which our Sun and its planets condensed many years later. Supernova explosions are highly relativistic events, converting something like one percent of the mass of a star into explosive energy, a much higher fraction than in a nuclear bomb. This spectacle of relativity was an essential step on the road to creating life on Earth.

It is in astronomy that the most spectacular consequences of relativity are found. We see protons (cosmic rays) that hit the upper atmosphere of the Earth traveling at speeds incredibly close to c . We see jets of gas shooting out of quasars at nearly the speed of light. We see spinning stars (pulsars) that rotate so fast that their surfaces are moving at one-tenth of the speed of light. We see regions of space containing black holes, that trap light and therefore everything else. In all of these phenomena, gravity also plays a key role in making them happen. It is time, therefore, to make our next step forward, to learn about general relativity. When we make the union of gravity with special relativity, we will begin to understand the Universe.

Spacetime geometry: finding out what is *not* relative

When Einstein began to develop his theory of gravity, he knew he had to build on special relativity, but he felt strongly that he also had to preserve Galileo's other great contribution to physics, the principle of equivalence (Chapter 1). As with special relativity, Einstein worked by blending the old and the new in equal proportions: special relativity combined the old principle of relativity with the new principle of the universality of the speed of light; in his new theory of gravity Einstein combined the old principle of equivalence with his new theory of special relativity.

Einstein required more than ten years, including six of intensive work, to bring these two principles together in a way that was also consistent with Newton's theory of gravity and with all the observational evidence. The resulting theory came to be called general relativity. Conceptually elegant but mathematically complex, it made a great number of new predictions, almost all of which are now verified by experiment or astronomical observation. General relativity turned Einstein into a household name, and justly so: it is one of the triumphs of theoretical physics.

The observational evidence that Einstein used was mainly the fact that Newtonian gravity was so successful in describing the motion of the planets. The one unexplained gravitational effect was the extra shift of the perihelion of Mercury's orbit, which we described in Chapter 5. Although Einstein knew about this problem, he did not use it to guide his development of general relativity; rather, he kept it to one side and used it as a test of the validity of his equations once he had arrived at them. As we describe in the next chapter, Chapter 18, his theory passed this test with flying colors.

Gravity in general relativity is ...

Let us repeat here the astonishing statement in the last paragraph: Einstein began his quest for a relativistic theory of gravity using essentially the *same* observational evidence about gravity that was available to Newton! The invention of general relativity was not driven by an urgent need to explain new experimental results. Einstein did have something that Newton did not, but it was a theory, not an observation: special relativity. Einstein's main objective was to achieve *theoretical consistency* between gravity and the rest of known physics. It is perhaps all the more amazing, therefore, that in the end Einstein devised a theory that made many new and completely unexpected predictions that could be tested by experiment and astronomical observation.

Our purpose for the rest of this book is to learn about general relativity and its applications. This will take us on a journey to some of the most interesting phenomena in astronomy. We will have to steer a careful course between the rocky shoals of too much mathematical complexity and the becalmed waters of over-simplification. There is a huge amount that can be understood well with the level of mathematics we use in this book, and readers will find that the phenomenology of relativistic

In this chapter: we take our first steps toward understanding general relativity by describing special relativity in terms of the geometry of four-dimensional spacetime. This geometry describes in an elegant and visual way the algebraic predictions of special relativity that we met in the previous chapters. The geometry of special relativity is flat, and we learn how the equivalence principle will allow us to curve it up and produce gravity.

>Underneath the text on this page is the familiar Mercator projection map of the entire Earth. This map illustrates strikingly the fact that the surface of the Earth cannot be represented faithfully on flat paper. The Earth is curved, and mapping it flat distorts distances. In this case, the distances near the poles are exaggeratedly large.

In this section: we look ahead at the ways we will learn to use general relativity in the rest of this book.

gravity can be understood, not just learned about, from the few basic principles that we develop, carefully, in this and the next two chapters.

Here are some of the things we will learn how to do.

- ▷ Chapter 18
 - We shall learn how to reproduce the effects of a Newtonian gravitational field by using Einstein's geometric ideas.
- ▷ Chapter 18
 - We shall see how to work out the gravitational deflection of light, getting the correct relativistic value instead of the Newtonian one we found in Chapter 4.
- ▷ Chapter 21
 - We shall compute the orbit of a planet around a black hole, and show that the orbit is not a closed ellipse but rather a precessing ellipse, describing a rosette pattern over time.
- ▷ Chapter 19
 - We shall learn that the main differences between the predictions of general relativity and Newtonian gravity can be traced to a difference in the *source* of gravity, and in particular the way that pressure helps to create Einstein's gravity.
- ▷ Chapter 19
 - We shall deduce that rotating stars and black holes must produce gravitational accelerations that resemble the magnetic forces of electromagnetism, in that they depend on the *velocity* of the object being accelerated.
- ▷ Chapter 20
 - We shall compute the structure of a neutron star and see why stars that are too heavy must collapse to black holes.
- ▷ Chapter 22
 - We shall compute the effect of a gravitational wave on a detector, and so understand why the new gravitational wave astronomy is so interesting.
- ▷ Chapter 23
 - We shall see how gravity creates some of the most beautiful pictures in astronomy, multiplying and distorting images of distant galaxies and quasars as they pass through gravitational lenses.
- ▷ Chapter 25
 - We shall calculate the history of our expanding Universe back to the Big Bang, learn how the elements hydrogen and helium were made, and speculate on how the huge amount of dark matter in the Universe helped stars and galaxies to form.
- ▷ Chapter 27
 - We shall understand, from the way pressure creates Einstein's gravity, why cosmologists believe that the Universe underwent a period of very rapid expansion at the beginning, and why its expansion may even today be accelerating rather than slowing down.
- ▷ Chapter 27
 - We shall glimpse the links between gravitation theory and the theories of the other fundamental forces in physics, as some of the brightest theorists working in physics today struggle to produce a theory of physics containing all the forces in one unified whole.

This is a tantalizing menu for the remainder of our exploration of gravity, but it also an indication of the broad sweep of applications of general relativity in astronomy today. Einstein's invention, devised purely for mathematical consistency, has become essential for the interpretation of the world we see around us. Gravity, the same everyday gravity that Galileo probed with his inclined planes, is the key to understanding the modern Universe.

These predictions of general relativity are radical enough, but what is even more revolutionary about the theory is the *way* it describes gravity.

Until Einstein, gravity was thought of as simply a force, like the electric force. Einstein described gravity instead as geometry.

Rather than being a force exerted by one body directly on another, gravity was more indirect: one body would cause space and time to curve, and the other body would move in response to this **curvature**. This is unfamiliar language for us: we are used to the idea of a force, but what does it mean that gravity is geometry? The purpose of this chapter and the next is to help us to understand Einstein's way of thinking about gravity.

... *geometry*

Since Einstein describes gravity in terms of geometry, our natural first question is, what do we mean by the word *geometry*? Consider ordinary spaces we are familiar with, such as the surfaces of spheres, or the **Euclidean plane** as represented by a flat piece of paper. All such spaces are smooth and continuous, but when we speak of their geometry we mean something more: we mean their shape, the distances between points in the space, and so on. We calculate distances typically by using coordinates. For example, if I give you the latitude and longitude of both New York and London, you could in principle calculate the distance between them along a great circle on the Earth's surface. This sort of calculation is routine for airlines.

Now, the latitude and longitude of a city are coordinates that locate it on the Earth, just as the x - and y -coordinates locate points on a graph. We generally need coordinates in order to specify which points (cities) we are talking about, and then we use them to compute the distance. But we know that the distance is something that does not depend on the coordinate system we use. For example, we might use longitude measured, not from the Greenwich meridian, but from (say) a line passing through Disneyland California: we could call this the Mickey Mouse coordinate system for the Earth. Although this would change the values of the longitude coordinate we use to describe every city, it would not change the distances between cities.

We want to describe the geometry of relativity. We have already seen that time and space must both be involved, since both are distorted and even mixed by the Lorentz–Fitzgerald transformation. We must therefore explore the geometry of spacetime, the four-dimensional continuum with three spatial dimensions and one time dimension that is the arena for special and general relativity. The unification of space and time into spacetime is one of the most important conceptual advances that special relativity led physicists to. We define and explore it in the next section.

The geometry of a space, like the Earth's surface, is described by the distances between places, not the coordinates of the places. It is something that is a property only of the space itself. When we study the geometry of special relativity and then of spacetimes with gravity, we will of course have to use coordinates (such as t , x , y , and z) to describe events in the spacetime. But we have seen that in special relativity two different observers will use different coordinates. The geometry of the spacetime must not depend on which observer describes it. So we must find ways of describing the geometry using invariant distances between events.

This invariant will be called the **spacetime-interval**. This is a word we have used often in this book to represent a particular lapse of time. In relativity it is used in a very specific manner, to represent a measure of separation of events in time or space that is agreed by all experimenters, independent of the coordinates of the events. We will define it later in this chapter and then use it repeatedly through the rest of the book. The geometry of spacetime is determined by the spacetime-intervals between events. Spacetimes that describe gravitational fields

In this section: we learn what geometry is and why it can be used to explain gravity. The key is a distance measure in spacetime called the interval.

will differ from the spacetime of special relativity by having different spacetime-intervals between events.

Spacetime: time and space are inseparable

In this section: spacetime is the four-dimensional arena in which all things can be described. Einstein showed that we cannot separate space from time easily. We learn the language of spacetime and illustrate the entanglement of space and time with an example loosely based on the legend of William Tell.

We will see how to describe the geometry even of special relativity.

Our study of special relativity in the previous chapter has already told us that the world is not constructed in the way we may have thought, and certainly not in the way that Galileo and Newton thought. Time is not absolute: different experimenters measure it differently, and no single experimenter has a better definition than another. Nor is space absolute: solid objects have different lengths when measured by experimenters moving at different speeds.

These ideas were just as troubling and counter-intuitive to physicists of Einstein's day as they are today to new students who encounter them for the first time. When physicists began to think more deeply about *why* the ideas were troubling, they found that it helped them to stop thinking about time and space as distinct and separate things, and instead to join them together. Since time is one-dimensional (the history of ancient Rome, for instance, can be ordered along a single line) and space has three dimensions, their combination is a four-dimensional realm. We call this spacetime.

A single point of spacetime occupies, therefore, both a particular location in space and a particular moment in time. Just as space is the collection of all "places", or *points*, spacetime is the collection of all "happenings", or events. A **spacetime diagram** is a graph that records the entire history of an experiment or of some other process.

We can clarify what this means by drawing a spacetime diagram, as in Figure 17.1. I will illustrate the idea by recording in this diagram, in a simplified way, the history of the legendary episode where the Swiss patriot William Tell was compelled to shoot an apple from the head of his son. The diagram can only show two of the four spacetime dimensions, so I have chosen to show time (vertically) and the x -direction of space (horizontally). I align my x -direction with the direction the arrow took when flying from Tell to his son.

In the left-hand panel of this figure, we imagine that Tell stands at the origin of the space coordinates ($x = 0$) and fires the arrow at time $t = 0$. The *event* of firing the arrow is the intersection of the time and space axes. The arrow does not remain at the origin, but instead moves to the right (positive x) as time increases. This motion is shown as a slanting dashed line. This is the history of the arrow's progress from Tell to the apple: at any time t one can find out where the arrow was by just looking at the point on the line where its t -coordinate value has the desired value. Such a line is called a **world line**.

Similarly, the apple has a world line. This is the vertical dashed line. It is vertical because the apple does not move while the arrow approaches: it stays at the same value of x all the time. The single *event* at which the arrow pierces the apple is then the intersection of the two world lines, shown as the gray dot. This event belongs to the histories of both the arrow and the apple. We see that *events* are represented as single points in the diagram: they have no extent in time or space. *Objects*, on the other hand, remain objects through time, so they are represented by continuous lines that go upward in time.

Notice that William Tell's son is not shown. Fortunately for him, his head's world line does not intersect that of the arrow! His head stayed at the same x -location as the apple, but it was at a different height (say, a different value of the coordinate z). Therefore we can't show it on this diagram: we have no room for the z -dimension here. Nor can we show Tell's subsequent escape from the scene, in which he rode off in the y -direction!

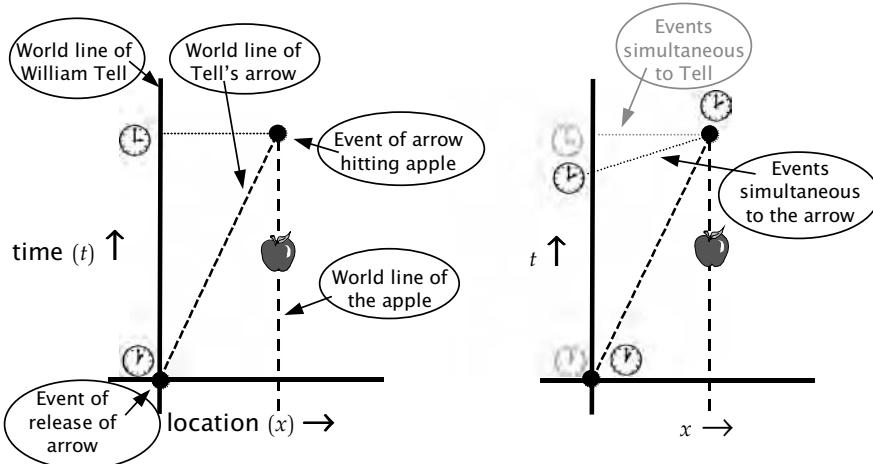


Figure 17.1. Spacetime diagram of what happened when William Tell shot the apple. The arrow starts at the spatial origin ($x = 0$) and travels to the right as time increases (slanting dashed line). The apple stays in one place (vertical dashed line). The intersection (gray dot) of the world lines of the arrow and apple represents the event at which the apple was pierced by the arrow. For an explanation of the way the diagram illustrates time dilation, see the text.

Relativity of time in the spacetime diagram

The idea of drawing this diagram might have occurred to Newton as easily as to us, so where does relativity come in? One way to introduce relativity is to ask about how much time it took the arrow to fly from Tell to the apple. Suppose, for example, that Tell was wearing a very accurate Swiss wristwatch, whose world line is the vertical axis because the watch (and Tell) stayed at the origin $x = 0$ during the flight of the arrow. Suppose, next, that Tell looked at his watch just when the arrow hit the apple. Importantly, we don't mean that Tell looked at his watch when he saw the arrow hit the apple: this would be later than the time at which the arrow actually hit, since it would take light some time to reach Tell so that he could see it. As a good experimenter, Tell would have to correct for the time it takes him to get the information that the arrow reached its goal. We assume that he does this, that he manages to look at his watch at exactly the moment (according to his measurements) when the arrow hits the apple. What time was it on the watch then? Since the two events did not occur at the same place (the watch was on Tell's wrist when he looked at it, while the arrow was somewhere else), relativity tells us we should be careful in answering this question.

Before we answer, let us look more carefully at the diagram itself. Who is the experimenter who recorded the time and position of the apple and arrow in order to draw the diagram? Assuming the experimenter was careful and accurate, the only important question is, what was his or her state of motion? Clearly, since the apple stays at the same x -position in this diagram, the experimenter was at rest with respect to the apple, and therefore at rest with respect to the ground. The diagram we have drawn is the natural way for such an experimenter (Tell himself, for example) to describe these events.

Tell's wristwatch is therefore a good recorder of this experimenter's definition of time. Since all events that occur at the same time are at the same height in this diagram, we only need to draw a horizontal line from the gray dot to the vertical time-axis in order to discover what time Tell thinks it is when the arrow strikes the apple. This line is shown as a dotted horizontal line in the left-hand panel.

But we know that a different experimenter, say one who was moving at the same speed as the arrow, would give a different answer to this. If Tell, in a demonstration of supreme self-confidence, had attached a similar Swiss watch to the arrow just before releasing it, then because of the time dilation effect, the flying clock would

In this section: we continue to work with William Tell, seeing here how to represent the effects of special relativity on time.

have ticked a little less time when it reached the apple than Tell would have seen when he looked at his wristwatch. Of course, for realistic arrow speeds, this would be an incredibly tiny time difference. Let's agree, for the fun of the story, that the watch is accurate enough to measure this difference!

Because the time on the flying watch was less than the time at which Tell looked at his own watch, the flying experimenter would have decided that the arrow hit the apple a little *earlier* than Tell believed: this experimenter believes that Tell looked at his watch too late. Put another way, the set of all events that the moving experimenter regards as being simultaneous with the piercing of the apple is a line that is not horizontal in this diagram: it tilts a little upwards to the right in order to intersect Tell's wristwatch's world line at a time that is earlier than the moment when Tell looks at the watch. This tilted line is the dotted line shown in the right-hand panel of the figure.

The two panels in this figure show, in a striking and graphical way, how the idea of time changed from Newton to Einstein. For Newton, time was universal: provided two experimenters synchronized their clocks and agreed to start them at the same time, they would always agree on what events occurred at what times. The left-hand panel of Figure 17.1 on the preceding page would represent Newton's concept of time correctly for any experimenter. Horizontal lines would connect events that were simultaneous, and all experimenters would agree on this.

But Einstein's time dilation, which leads to the loss of simultaneity, changes all that. For the experimenter at rest with respect to the apple, lines that are horizontal in this diagram are lines of constant time. For the experimenter flying with the arrow, lines of constant time are tilted with respect to the first set of lines.

The flying experimenter would of course not draw them tilted in his or her own spacetime diagram. In that diagram the arrow would be at rest, the moving apple would follow a world line slanted to the left, and Tell's line of constant time would be tilted upwards to the left. Readers may find it interesting to draw a spacetime diagram the way the flying experimenter would.

Time dethroned ...

In this section: the special role that time played in Galilean and Newtonian physics does not continue in relativity. Instead, we need to know about space and time together.

If time is not universal, do we have to forget completely our present notions of time? No, because relativity preserves the most important aspect of time, which is the separation of cause from effect. Relativistic time keeps a consistent direction: we saw in the previous chapter that the ideas of future and past are still well-defined. Different experimenters may disagree on the order in time of events that cannot have a causal relationship with one another, such as the event of piercing the apple and the event of Tell glancing at his watch. But they will always agree on causally related events, such as the fact that Tell fired the arrow *before* it hit the apple.

In relativity, however, there is not just one time. Time is best regarded as a coordinate in spacetime, just as the value of x is a coordinate. We know what it means that x is "just" a coordinate: two different experimenters could orient their x -axes in different directions. For example, there was no need for me to have put the flight of Tell's arrow on the x -axis; the y -axis or halfway between would have done just as well. We are used to thinking that there is no special, or preferred, orientation for the x -, y -, and z -axes. Since Einstein, time is the same kind of thing: one experimenter draws the line of simultaneity horizontally in the spacetime diagram in Figure 17.1 on the previous page, while another draws it slightly tilted.

This tilting of the lines of constant time is the main reason that physicists find it so useful to think of the four-dimensional spacetime as a single continuous entity. Three-dimensional space could be represented by a horizontal slice through spacetime, just as we represent the x -axis by a horizontal line in the figure. But how

horizontal: which experimenter do we use to define the slice? There is no unique way of identifying three-dimensional space within spacetime. Spacetime is a single continuum; space and time can be separated from one another only by choosing the coordinates of a single experimenter, and the separation will depend on which experimenter we choose.

So far, we have only described spacetime as a kind of four-dimensional map that charts not simply places but entire histories. This is how we will look at the Universe through the rest of the book. But to describe gravitation, we need to put some real geometry into spacetime.

It is not enough just to have a map. We have to know whether the map is flat, curved, crumpled, perforated, whatever: gravity is in the crinkles in the map of spacetime!

...and the metric reigns supreme!

The fundamental reason that relativity merges time and space into spacetime is that time and space are separately not invariant: different experimenters get different results for the length of time something takes, or the distance between them. But spacetime itself is invariant, and it is one of the most remarkable facts about special relativity that it provides us with a new, invariant, unified measure of distance in *spacetime* itself. This measure, which is called the spacetime-interval, is a combination of distances in space and in time. It measures the spacetime distance between *events*. And it is invariant: all experimenters who measure times and distances carefully will get the *same* value for the spacetime-interval between any two events, even if they get different individual values for the time difference and distance between the events.

Here is the definition of the spacetime-interval. Suppose, as measured by a certain experimenter, two events are separated by a time t and a spatial distance x . Then in terms of these numbers the spacetime-interval between the two events is the quantity

$$s^2 = x^2 - c^2 t^2. \quad (17.1)$$

Notice that this is written as the square of a number s . The spacetime-interval is the quantity s^2 , not s . In fact, we will not often deal with s itself. The reason is that s^2 is not always positive, unlike distance in space. If ct is larger than x in Equation 17.1 then s^2 will be negative. In order to avoid taking the square-root of a negative number, physicists usually just calculate s^2 and leave it at that. You should just regard s^2 as a single symbol, rather than as the square of something.

This quantity is important *because* it is invariant. Two different experimenters can calculate it and will get the same answer. Let us see how this happens by first calculating a spacetime-interval from William Tell's measurements, and then from those of the flying experimenter. We will compute the spacetime-interval between the following two events: the first event is the firing of the arrow, and the second is the piercing of the apple.

Suppose that, as measured by Tell, the distance of the shot was ℓ and the time of flight of the arrow was t . Then the spacetime-interval, as measured by Tell, is

$$(s^2)_{\text{Tell}} = \ell^2 - c^2 t^2.$$

Notice that this will come out to be negative. The distance light would travel during the flight of the arrow is ct , and this must of course be much larger than the actual distance the arrow traveled, ℓ . Therefore the spacetime-interval between these events is negative.

In this section: the rule for calculating intervals in spacetime is given by the metric. The intervals are experimentally measurable and will be agreed by all experimenters, just as distances computed from the theorem of Pythagoras are agreed by all measurers in Euclidean geometry. This is the key to an invariant idea of the geometry of spacetime.

►There are a number of different variations in the definition of the spacetime-interval. Some scientists multiply it by -1 or divide it by c^2 before calling it the spacetime-interval. Since these multipliers are constants, these differences are simply matters of convention, like measuring spatial distances in miles or kilometers.

▷ If this sounds strange, place yourself on a moving train going into a tunnel: from your perspective, first the entrance to the tunnel rushes over you with a whoosh, then there is the noise of the tunnel, and then the end of the tunnel passes over you and the sky reappears. You have not moved from your seat during all of this.

The flying experimenter, moving with the arrow, sees space and time differently. For one thing, the arrow stays in one place relative to this experimenter, so the distance x between the two events is zero. In this view, first William Tell and then the apple fly past the experimenter. Tell releases the arrow just as he passes our experimenter, the arrow stands still, and then the apple smashes into it! As we have seen above, the time τ between the two events, as measured by the flying experimenter (or equivalently by the Swiss watch attached to the arrow), is shorter than that which clocks on the ground measure. By the time dilation formula, the time τ is

$$\tau = t/\gamma.$$

From this we can easily compute the spacetime-interval measured by the flying experimenter, bearing in mind the definition $\gamma = (1 - v^2/c^2)^{-1/2}$. Since the distance is zero, we have simply

$$(s^2)_{\text{flier}} = -c^2 \tau^2 = -c^2 t^2 / \gamma^2 = -c^2 t^2 (1 - v^2/c^2)^{-1} = -c^2 t^2 + v^2 t^2.$$

Now, the product vt is just the distance ℓ that the arrow flies as measured by Tell, which shows that this spacetime-interval is exactly the same as the spacetime-interval as computed by Tell. Even though the two experimenters measure different values for distance and time, the differences are just what is needed to insure that the spacetime-interval they compute is the same for both.

The spacetime-interval between two events does not depend on the experimenter who defines time and space. A different experimenter may, because of Lorentz–Fitzgerald contraction and time dilation, assign different values to Δt and Δx , but the spacetime-interval between two given events, calculated using Equation 17.1 on the previous page, will be the same for all experimenters.

The spacetime-interval is the most fundamental number associated with pairs of events. It gives a distance measure on spacetime. You may already have noticed that it has a certain similarity to the Pythagorean theorem, which defines distances in the ordinary two-dimensional plane:

$$\ell^2 = x^2 + y^2. \quad (17.2)$$

The spacetime-interval differs from this equation by using a minus sign in front of one of the terms. This is an important difference. If two events occur at the same time (as measured by some experimenter) then the spacetime-interval is the same as the square of their Pythagorean distance, which is positive. But if they occur at the same place at different times then their spacetime-interval will be the negative of the square of the time between them. This change of sign is what keeps time distinct from space in relativity: although neither is absolute, and they will be separated from one another in different ways by different experimenters, time and space are not identical, and the different sign in the spacetime-interval is the way the distance measure on spacetime respects that difference.

The formula for the spacetime-interval also contains the constant c , which can be thought of as a weighting factor between space and time, telling us how much a given time difference contributes to the spacetime-interval in relation to distances. Since the weighting is a constant, it is not particularly important here. But we will see when we discuss curved spacetimes below that *curvature* comes about essentially when the relative weightings of terms in the spacetime-interval change from place to place in the spacetime. Gravity is represented mathematically by the weighting

factors in the spacetime-interval. Because the weighting factors in special relativity are constant, we say that the spacetime of special relativity is *flat*.

The spacetime-interval has another name: the **spacetime metric**. The word *metric* is used in geometry for distance measures, particularly those that involve the squares of coordinate separations. Our discussion of relativity shows us that the metric takes over the role that, in Newtonian mechanics, was shared by separate measures of time and distance.

The geometry of relativity

Now we are ready to talk about geometry: we have a space (called spacetime) and a distance measure (the spacetime-interval). Einstein himself did not at first seem to think geometrically about spacetime. It was his former mathematics professor in Switzerland, Hermann Minkowski (1864–1909), who pointed out how important the geometry of spacetime was. Because of his contribution, physicists refer to the spacetime of special relativity by the name **Minkowski spacetime**.

Figure 17.1 on page 215 is a depiction of part of Minkowski spacetime. I have not marked time and distance units along the axes. If I had, and if I had used conventional units like seconds for time and meters for length, then the world line of a photon would be so tilted over that one would not be able to distinguish it from a line parallel to the x -axis. Its slope, $\Delta t/\Delta x = 1/c \approx 3 \times 10^{-9} \text{ s m}^{-1}$, would be too shallow to draw. Put another way, in one second of time (the vertical direction in the diagram), a photon would travel such a large distance in x (the horizontal direction) far that it would not only be off the page, it would be off the Earth!

This would therefore not be a good way to draw a diagram of a part of spacetime in which we want to record relativistic effects. Since relativity only differs from Galilean and Newtonian physics when things move at speeds close to c , it is better to use different *units* in a spacetime diagram to keep a photon's world line at a reasonable angle. The units that many physicists use for spacetime diagrams involve a re-scaling of the time coordinate to one we shall call T :

$$T = ct. \quad (17.3)$$

This is shown in Figure 17.2. The time coordinate has been expanded so much by this re-scaling that now a photon world line has a slope of one. This new time coordinate has dimensions of distance: we measure time by the distance light travels in that time. One meter of time is the time it takes light to go one meter, or $3 \times 10^{-9} \text{ s}$. This is similar to something that most people are familiar with, measuring distances in light-years. One light-year is the distance light travels in a year. If we had re-scaled the x -axis to light-seconds, and kept time in seconds, we would similarly have produced a photon world line with a slope of one. But it is more conventional in relativity these days to re-scale time, so that time is measured in "light-meters".

In terms of these new units, we can write the spacetime-interval in the simpler way:

$$s^2 = x^2 - T^2. \quad (17.4)$$

The striking thing about this distance measure is that it can be either positive or negative. This is very different from the distance measure in ordinary space. This means that the geometry of spacetime will have something that is not familiar from ordinary **Euclidean geometry**: the spacetime-interval between two well-separated events can actually be zero. Let us face this squarely: what is a zero spacetime-interval?

In this section: the spacetime of special relativity is called Minkowski spacetime. Although there is no universal way to separate it into time and space, there are invariant divisions of spacetime. They are given by light-cones.

►When we learn about spacetimes that represent gravitational fields – black holes, cosmologies – we will see that many are named after their discoverers: Schwarzschild, Kerr, Friedmann, and so on. Minkowski has the distinction of having been the first to describe an important spacetime in relativity, the spacetime that has no gravitation!

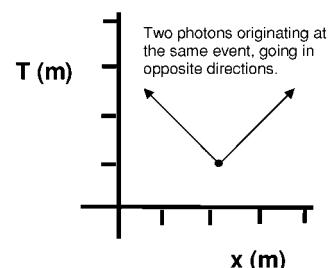


Figure 17.2. A spacetime diagram in natural units, showing the world lines of two photons, one traveling to the right and the other to the left.

If I set $s^2 = 0$ in the previous equation, I find $x = \pm T$. Remembering the definition of the re-scaled time T in Equation 17.3 on the preceding page, this implies

$$x = \pm ct. \quad (17.5)$$

This is just the equation for something moving at the speed of light. The positive sign is for a photon going to the right, the negative sign to the left. The two lines that are drawn in Figure 17.2 on the previous page are the lines that go through all the events that have zero spacetime-interval from the event at the origin of the diagram. We call these lines the **light-cone**, because of what it would look like if we added a further spatial dimension to the diagram. If we include the y -axis, say pointing out of the page, then there are world lines of light that move out from the origin in all directions in space, always moving forward in time at the speed of light. These lines, taken together, form a cone whose apex is at the origin. This is the light-cone of the origin.

Other events have light-cones too: the set of all light world lines that pass through a given event is the light-cone of that event. Any event on this light-cone will have a zero spacetime-interval from the original event. Since the spacetime-interval is independent of which experimenter measures it, the light-cone of any event is an *invariant*: all experimenters will assign the same events to the light-cone of any given event. If you think about this, you will see that this is nothing more than one of the fundamental principles of special relativity: all experimenters measure the speed of light to be the same value c .

Events have light-cones going into the past as well. These consist of all light world lines that converge on the given event from the past. We speak of the past light-cone and the future light-cone of any event.

The invariance of the light-cone has other consequences. It divides spacetime into separate regions, relative to a given event. The interior of the future light-cone consists of events that are separated from the given event more by time than by space, so they have negative values of their spacetime-interval from the event. They are called the timelike future of the event. Similarly, the timelike past is the interior of the past light-cone. The exterior of the light-cone is a single region whose events are separated from the given event (the one at the apex of the cone) more by space than by time, so this region is the spacelike “elsewhere” of the given event. All experimenters will agree on this division of spacetime relative to a given event.

Proper measures of time and distance

In this section: the spacetime-intervals lead to definitions of proper time and proper distance that all experimenters will agree on.

Just as the Pythagorean distance in space is the true distance that someone would measure if they walked along the line, so is the spacetime-interval a measure of the true distance, or **proper distance** in spacetime. If I want to measure the length of something, even say a moving train, then as we saw in the last chapter I must make the measurement at a given time: I have to take the distance between the locations of the ends at the same time, according to my own clocks. This means that when I compute the spacetime-interval between the events that I used (the events that correspond to the locations of the two ends at the given measurement time), then the time-difference is zero and the spacetime-interval will be exactly the square of the distance that I measure. We say that the spacetime-interval gives the proper distance between two events that have a spacelike separation; in other words, it is the distance that an experimenter would measure between them if the events were simultaneous to the experimenter.

The same holds for timelike spacetime-intervals. For example, we saw above that the watch that William Tell attached to the arrow ticked a time whose square was just (in our units) the negative of the spacetime-interval between the events. So the

spacetime-interval along a timelike world line tells us what the rate of ticking is of a clock that moves along that world line. All we need to do is to take the square-root of its absolute value, and divide by c if we want time measured in time units instead of distance units. This measure of time, the ticking of a clock that moves along a world line, is called the **proper time** along the world line, and is usually denoted by τ :

$$\tau = \frac{1}{c} \left| s^2 \right|^{1/2}. \quad (17.6)$$

It is interesting to ask what are the events in spacetime that have a given proper distance or proper time from the origin, say in Figure 17.2 on page 219. In ordinary space, the points that are at the same distance from the origin form a circle (in two dimensions) or a sphere (in three). What is the analog in spacetime?

In Minkowski spacetime, the points at the same *spacetime-interval* from the origin satisfy an equation of the form

$$x^2 - T^2 = \text{const.} \quad (17.7)$$

This is the equation of a hyperbola in the $x-T$ space. We call this the **invariant hyperbola**.

One such hyperbola is shown in the diagram in Figure 17.3. Since the spacetime-interval tells us the time measured by a clock traveling on a world line (the proper time), a clock that moves on a straight world line from the origin to any point on the hyperbola ticks the same total proper time, regardless of which world line it moved on. Since different straight world lines describe clocks moving at different speeds, the points on the hyperbola are the events that can be reached from the origin in a fixed given proper time (in the case shown, this is 5 m of time, or 1.6×10^{-8} s) by traveling at different speeds.

Because of time dilation, the clocks that travel faster take longer to tick out the given proper time, so one can see in Figure 17.3 that they go further than one would have expected without special relativity. As an extreme case, it would be possible in principle to send a team of astronauts across the Galaxy and back in a proper time as short as, say, a year, if we could accelerate their rocket to a speed sufficiently close to c . This was the basis of the time dilation fantasy at the end of the last chapter.

Equivalence principle: the road to curvature ...

The spacetime of special relativity is the simplest one to describe, which is why we have spent half of this chapter on it. But it is also the most boring of all the spacetimes we will meet: it has no gravity, it is flat and the same everywhere, it is just a static background arena for the events that happen in it. Spacetimes that have gravity are more interesting: they participate in the physics that happens in them, they affect the physical systems that they contain, and they make possible all kinds of new phenomena, such as black holes (Chapter 21) and gravitational waves (Chapter 22). They do this through their curvature.

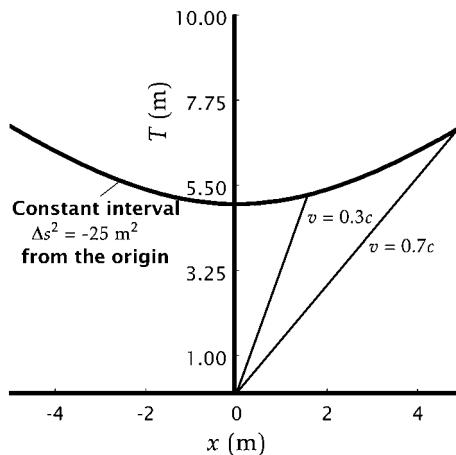


Figure 17.3. The hyperbola connects all the points that can be reached from the origin in a proper time of 5 m (measuring time in light-meters) by clocks traveling at different speeds. Two such clock world lines are shown.

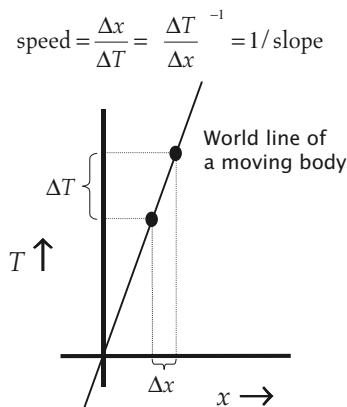
In this section: the spacetime of special relativity is flat. Curvature is needed to describe gravity. The key to understanding what curvature means in spacetime is the equivalence principle.

It is the equivalence principle that nudges us along the road to the picture of gravity as curvature of spacetime. According to Galileo's version of the principle of equivalence (Chapter 1), the effect of gravity on a body depends only on its state of motion: given an initial position and velocity, the subsequent motion of the body is fully determined. The body's color, number of baryons, charge, and so on do not affect its gravitational acceleration.

Now, the location and velocity of a body are both properties of the body's world line. Consider the world line of a body going at a constant speed in the x -direction, shown in Figure 17.4. The speed can be inferred from the slope of the line. This would also be true of an accelerated body, whose world line would not be straight: its slope at any point gives its speed at that time.

▷ This means the slope of its tangent at that point.

Figure 17.4. The world line of a body contains the information needed to compute its speed. Unlike the path in ordinary space, the world line tells us where the body is at any time, so one can read off its speed. The diagram shows that the speed between two events is the inverse of the slope of the world line joining the events. Since the diagram uses the time coordinate $T = ct$, the inverse of the slope is the body's speed relative to c , in other words v/c .



Now, according to the equivalence principle, if we know the location and the velocity (the speed and its direction) of a body, then gravity completely determines what its subsequent motion will be. Put another way, if we know that the world line of a body passes through a certain event with a certain slope, then that information completely determines the rest of the body's world line, in a given gravitational field. Every particle that starts out from that event with that slope (velocity) will follow the same world line, provided of course that gravity is the only influence on it.

This means that gravity determines *preferred* world lines, the ones which all particles follow, regardless of their mass, color, etc. The world lines of particles as they move through

spacetime under the influence of gravity can therefore be thought of as properties of the spacetime itself, not of the particles: it is irrelevant whether it is a proton or a piano in orbit around the Sun, the orbit will be the same. Einstein reasoned that we should focus our attention on these world lines, and find a description of gravity that showed why certain world lines were special.

Notice that it is crucial that we be in *spacetime* for this geometrical argument to work: it would not work in ordinary three-dimensional space. The path of a body in ordinary space traces out where it has been, but does not tell us when it was there, so it does not tell us its speed. William Tell's arrow, once released, follows the world line of a free body until it encounters the apple. But its spatial path, which we have drawn as a portion of the x -axis, could equally well be followed by a bird, supported by the aerodynamic forces on its wings. The spatial path by itself does not contain enough information to tell us about the effect of gravity. There is thus no possibility that gravity could be related just to the geometry of ordinary space. The equivalence principle tells us to think about the geometry of the full *spacetime*.

... is a geodesic

In this section: freely falling bodies follow world lines in spacetime that are as straight as possible. These are called geodesics. Orbita are curved lines when projected just into ordinary three-dimensional space, but in spacetime they are locally straight.

Now, mathematicians know that curved spaces have special sets of lines. Consider one of the simplest examples of a curved space, the ordinary sphere. The special lines on the sphere are the great circles. These are the curves of "least effort": if you walk on a large sphere (like an idealized Earth), and if you just keep walking straight ahead and never bother to make up your mind to change direction, then you will walk along a great circle.

Suppose you meet a little dimple in the sphere, a shallow basin rather like a

crater on the Moon. If you keep “following your nose” in this way, your path will generally emerge from the dimple going along a different great circle. Your path has been deflected, but you never decided to change your direction. You followed your nose, putting one foot in front of the other, and the geometry decided your direction for you.

By following your nose, you are imitating Newton’s first law of motion: once a body is set in motion, it will continue on a straight line unless acted upon by a force. Einstein’s new idea was that gravity should *not* count as an external force acting on the body, but rather that bodies affected by gravity are really just obeying Newton’s first law. They keep going as straight as they can. If their world lines curve, that is because spacetime itself is curved.

These special paths are called **geodesics**. Geodesics are defined as lines that go as straight as possible on a surface.

If the geodesics of a surface are straight, so that geodesics that start out parallel remain parallel and don’t intersect one another, then mathematicians call the surface a **flat space**.

In the spacetime of special relativity, particles move on straight world lines. The spacetime of special relativity, Minkowski spacetime, is therefore by this definition a flat spacetime. So this definition of flat and curved is consistent with our earlier, rather vaguer, notion that Minkowski spacetime is flat because it is boring!

The equivalence principle: spacetime is smooth

Everyone has experience of curvature by handling curved surfaces, so it is helpful to try to understand geodesics on such surfaces first. One of the first distinctions we make about surfaces is whether they are smooth or have a corner. A curved surface that is smooth can still bend sharply, but it does not have an edge where a lot of bending occurs all at once. The difference is the following. If you look closely at any point on a smooth curved surface, then in a small enough area around it, the surface is effectively flat: it is not much different from a flat piece of paper. If you look closely at a point on a true corner, it is still a corner: there is always the same amount of bending, no matter how small an area around the point we look at.

In this section: the equivalence principle links the spacetime of special relativity with that of general relativity: curvature is noticeable only over large regions.

In general relativity, we always assume spacetime is smooth in the same way that a curved surface is smooth. Gravity may be strong, but it does not concentrate its curvature in sudden jumps.

Now, consider the geometry of a very small patch of a smooth surface. If we stay near enough to the central point, it is hard to tell that the surface is curved at all. Think, for example, about the Earth. On a large scale, the Earth is curved: we all know that it is impossible to draw a map of the entire Earth on a flat piece of paper and expect it to represent faithfully the distances between all the points. Maps of the Earth are generally cut in a number of places to keep their distortions to a minimum, and even these maps are not perfect. However, maps of cities do not have such problems. To the accuracy with which we need to represent the distances between places in a city, the curvature of the Earth is not important. Of course, local curvature caused by hills might still prevent a map from being faithful. In that case, faithful flat maps could only be made for smaller areas, such as a section of the hillside. Mathematicians say that the Earth (and any other smoothly curved body) is **locally flat** because we can draw a local map that is flat and is as accurate as we want it to be, provided we cover a small enough part of the space with it.

▷This is almost always true, but like most rules it can have exceptions. The cosmic strings that we will discuss in Chapter 19 are exceptions.

For gravity, the local flatness of spacetime means that, in a small part of spacetime, gravity does not matter: spacetime looks like a portion of Minkowski spacetime. We have actually met this before, in Chapter 2. There we saw that the modern way of phrasing the equivalence principle is that a freely-falling experimenter does not see any effects of gravity, at least in a small enough region. We now see that we can re-phrase this in geometrical language:

The curvature of spacetime is smooth, and the locally flat observers are the freely-falling experimenters. Experimenters who fall freely with exactly the Newtonian acceleration of gravity see no gravity; photons and other particles move on straight lines in their local coordinates. The equivalence principle tells us that the geodesics of a spacetime with gravity are the paths of freely-falling objects.

Local flatness gives us a way of drawing a geodesic on a curved surface, at least in principle. If you are drawing a great circle on the sphere and then hit that dimple, how should you continue? Just make a small map of the locally flat region around your present location, near the (smooth) dip into the dimple. Make the region covered by the map much smaller than the dimple. Then follow the straight line on the map that is going in the same direction you have been going up to now. When you get to the edge of this map, draw a new one, also very small, around the new location and do the same thing. Always go straight according to the map. In the end you will have changed direction, because all the little maps can't be joined into one large flat map. But it is by following the little maps that you can determine your geodesic path.

Einstein's great insight was to understand that the equivalence principle lead naturally to a picture in which the effects of gravity could be represented by a curved spacetime whose geodesics, constructed in the way we have just described, are the paths that particles follow in the gravitational field. What is more, two locally straight paths that start out near one another will not remain exactly parallel; they may diverge or cross as they move through regions of slightly different curvature. This is a perfect representation of what we called "tidal forces" in Chapter 5: the forces that make nearby freely-falling particles diverge or converge. So the curvature of spacetime describes the tidal forces. Since the tidal forces are the part of gravity that can't be removed by going to a freely-falling observer, the curvature of spacetime *is* gravitation.

Einstein came to his insight early in his search for the right theory of gravity, and in fact there is no reason why other scientists could not have come to the same conclusion much earlier. After all, there is nothing relativistic about this picture of gravity as curvature. In Newtonian physics there is also a spacetime, which Newton did not talk about, but which we drew in Figure 17.1 on page 215. Newtonian gravity incorporates the equivalence principle, so it would have been possible for anyone who knew geometry to have reformulated Newtonian gravitational mechanics in this way. Scientists have done this since, but nobody did this before Einstein.

Since we are comfortable with Newtonian gravity, the best way for us to get a feeling of what gravitational curvature means is to find a curved-spacetime description of Newton's theory. We want to blend the equivalence principle with special relativity, as Einstein did. In the next chapter we will start with the spacetime of special relativity and follow what we have learned here to add in enough to get a spacetime with Newtonian gravity. This will enable us to get a good idea of what Einsteinian gravity is all about.

Einstein's gravity: Einstein climbs onto Newton's shoulders

Geometry is at the heart of Einstein's picture of gravity. The best place to see how gravity as curvature works is in the Solar System, where the predictions must be very close to the description given by Newton. In this familiar arena, we can compare the old and new ways of looking at gravity. In this arena, too, general relativity meets and passes its first two crucial tests: explaining the anomalous advance in the perihelion of the planet Mercury (which we puzzled over in Chapter 5), and predicting that light should be deflected as it passes the Sun by twice the amount that would be calculated from Newtonian gravity (see Chapter 4).

We saw in the last chapter that the equivalence principle tells us that it is not possible to represent gravity just by the curvature of space; the curvature of spacetime must include the curvature of time as well.

At this point, you may ask (indeed, you *should* ask) "What does curved time *mean*? What does it *look like*?"

Curved time sounds at first like a formidably abstract idea. But it is not nearly as abstract as it may seem. In fact, we will see below that we already know it by a different name: Newtonian gravity. The link is the way that gravity affects time, the gravitational redshift. The curvature of time is just the fact that the gravitational redshift is different in different places, i.e. that the gravitational field is not uniform. The way this leads to Newtonian gravity is already contained in our earlier discussions of the equivalence principle and of the deflection of light as it passes the Sun: we will just have to look at those discussions in a new way below in order to see Einstein's gravity in its simplest form.

Since the curvature of space must be a new form of gravity. It was the first really new element that Einstein added to gravity in the Solar System. We shall see how space curvature makes the new predictions that we mentioned above and that established the correctness of Einstein's theory, the deflection of light by the Sun and the precession of the perihelion of Mercury.

Our first step towards relativistic gravity must be to learn how to describe a curved spacetime. The best way to do that is first to describe a curved surface, such as the surface of a sphere, or of a more irregular object. The main idea always is to describe distances: you know what a surface is like if you can calculate distances along it.

Driving from Atlanta to Alaska, or from Cape Town to Cairo

A good way to develop an understanding of how to measure distance on a curved surface is to think about driving a car on a long journey. To guide you on this trip you need maps. Maps come in a variety of kinds: you can get large-scale maps of whole countries, regions, even continents; or you can get fine-scale city maps showing each small street.

In this chapter: we use Einstein's geometrical picture of gravity to study the motion of planets and light in the Solar System. We learn how to understand the curvature of time, and why Newtonian gravity is fully described by this curvature. We work out how the curvature of space changes the Newtonian deflection of light and makes Mercury's orbit precess. Since the extra deflection of light has been measured, we know what Solar System curvature Einstein's equations must predict when we encounter them in the next chapter.

>Under the text on this page is an image of Mercury, the only astronomical object known in the nineteenth century whose motion could not be explained by Newtonian gravity. Einstein's theory, constructed without reference to Mercury's motion and without any freedom to adjust the theory to explain the motion, nevertheless exactly predicted the anomalous extra motions that had been observed. This triumph convinced Einstein and many other astronomers that general relativity had to be correct. The image is a photomosaic recently processed from images taken by the Mariner 10 mission. Courtesy NASA and the Astrogeology Team, US Geological Survey.

In this section: coordinates are familiar to anyone who has navigated by using maps. Estimating distances on a map requires re-scaling the map reference coordinates.

Figure 18.1. This map of the area around New York City in 1970 used grid lines spaced by half a degree in latitude and longitude. Since this is not near the equator, these steps have different lengths. To calculate distances from references to these grid lines, one would have to use the weighted form of the Pythagorean law, Equation 18.1. The weighting factors are given in fact by Equation 18.2. Image from the U.S. National Atlas, courtesy of The General Libraries, The University of Texas at Austin.



The city maps are usually drawn as if the city were flat and the geometry were Euclidean, and they often place North at the top of the map. A good map will contain reference numbers, such as squares labeled by numbers going across the map and letters going down the side, so that a given region of the city is denoted by, say, reference 3E or 1P. These references are just *coordinates* for the map.

The map should give you a distance-scale, and you could use this to convert the coordinates to standard Cartesian coordinates, so that you could use a distance x instead of a location "1", y instead of "D", so that reference location 1D is the same as (x, y) . Then you could measure diagonal distances across the map using the Pythagorean rule given in Equation 17.2 on page 218,

$$\ell^2 = x^2 + y^2.$$

It could happen that the grid reference lines have different spacing in the two directions. That is, the width of a single reference block in the horizontal direction (going from 1 to 2, say) is a distance A in kilometers, while the height of a vertical reference block (from C to D) is B kilometers. An example of such a map is shown in Figure 18.1. To convert this back to an equation giving the true distance between map reference points one would have to use the distances A and B as weighting factors on the changes in the reference locations:

$$(\Delta\ell)^2 = A^2 \times (\text{change in horizontal reference})^2 + B^2 \times (\text{change in vertical reference})^2. \quad (18.1)$$

Notice we have introduced a subtle variation here: the second equation deals with *changes* in coordinates and the distance between them. We will find below that the best way to describe distances along surfaces is to take very small steps, using the Pythagorean rule in a form like Equation 18.1, and add up the steps. This equation allows us to calculate what we called proper distance in the last chapter: the true distance independent of coordinate system.

This is fine for getting around the city or its metropolitan area, but if your journey is long, you will soon leave the domain of this map. What then?

Suppose that all you have to guide you is a big collection of local maps of all the cities, towns, and rural regions. You go from one map to another as you drive. You could navigate this way, but you would forever be puzzling over the map edges. How does the road we took on the last map match up to one we are on now? Our

road left the edge of that map at map reference 1E; is it the same as the road entering the edge of the new map at 8A? How far did we travel to get across that last map? Is the scale on the new map the same? Is it oriented so that North is at the top?

This is messy, but it has the advantage that you are always looking at a Cartesian representation of distances. Once you get used to the scale of each map, you can quickly estimate distances from the Pythagorean rule as given in Equation 18.1. But you constantly have to work out how the maps join. There is a better way to navigate, and that is to buy a single map, drawn in a smooth way, with a single map reference system (a single coordinate system) that unifies the whole region through which you travel. You then use the smaller maps only if you need fine detail somewhere.

This makes navigation vastly simpler, but it introduces some distortions. For example, if the region through which you journey is large enough, you may find that the vertical map reference lines are only approximately North–South lines, and that the true northerly direction changes relative to these lines as you move across the map. If the journey takes you far from the equator, then the further you go the smaller are the East–West distances for a given change in the map reference: on a flat map, Canada and Scandinavia look much bigger than they actually are. (See the Mercator projection of the Earth on page 225 for an extreme example.) In other words, it is not possible to keep the *scale* of the map the same everywhere.

You can live with these distortions, of course, if you are aware of them. Suppose, on your large map covering the entire journey, you wanted to write down a rule for calculating distances anywhere on the map, in terms of the map's reference system. You want to extend Equation 18.1 to be valid everywhere, not just in one local map. The extension is not difficult. We just have to recognize that the numbers A and B convert from coordinates to real distances, and that this conversion may be different in different places as the scale changes gradually across the map. Your distance rule would look the same as Equation 18.1, but now A and B would be *functions of position*. Their values would depend on where you are on the map.

Dimpled and wiggly: describing any surface

Let us look at a concrete example of what we have just described. Suppose my map has reference coordinates that I call x and y , but which are *not* Cartesian: they just stand for whatever reference numbers go across and down the map, respectively. Suppose that I find that distances on my journey are well-described by the following form of the Pythagorean rule:

$$(\Delta\ell)^2 = (\Delta x)^2 + \sin^2 x (\Delta y)^2. \quad (18.2)$$

This is a special case of Equation 18.1, where we have re-named the coordinates (map references) and taken $A = 1$ and $B = \sin x$. Now, what surface am I describing? What does it look like?

Don't assume that, just because I have called the coordinates x and y , this is a flat plane! Even if we used "Martha" and "Fred" for the map reference coordinates, the geometry would be the same.

Take a guess at the answer for the geometry before you read the next paragraph.

The answer is that the space described by Equation 18.2 is actually the *surface of a sphere of radius 1*. It describes the entire Earth, with map reference coordinates that are the usual spherical polar coordinates: the coordinate x is the polar coordinate angle that is usually called θ and the coordinate y is the polar coordinate angle ϕ , both angles measured in radians (see Figure 18.2). Put another way, Equation 18.2

In this section: how to compute distances along a curved surface in any kind of coordinate system.

>Q: What is the longest distance you have ever driven, measured in Earth radii?

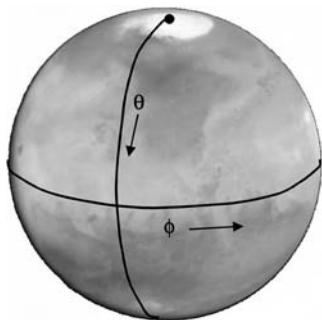


Figure 18.2. The way the latitude and longitude coordinates θ and ϕ run on a sphere, illustrated using a picture of Mars taken by the Hubble Space Telescope. (Photo courtesy NASA/HST/STSCI.)

gives the distance between nearby points on the Earth if we identify the coordinate y with the *longitude* and the coordinate x with the **co-latitude** (90° minus the latitude, so that the North Pole is at $x = 0^\circ$ and the South Pole at 180°). We also need to set the radius of the Earth to 1, so that means we are using a distance-scale in which one unit of distance equals 6400 km.

To see why this equation describes a sphere, consider latitude and longitude on the Earth. Forget Equation 18.2 on the preceding page for the moment. Just think about moving all the way around the Earth on a circle of constant latitude, i.e. keeping a constant distance from the Pole. How far have you gone? You can't answer this without knowing your latitude. A journey around the Earth along the equator is longer than a journey around the Earth at a high latitude.

Equation 18.2 answers this question mathematically. Remembering that x is co-latitude and y longitude, we see that the effect of the factor $\sin^2 x$ is to change a given difference in longitude Δy into the corresponding distance in the correct way. (Actually, of course, it changes the squared longitude difference into the correct squared distance.) The closer one goes to the pole, the smaller is the co-latitude x , the smaller is $\sin x$, and so the smaller is the actual distance associated with the journey. The factor of $\sin x$ is just the radius of the circle of constant co-latitude x , as measured in units of the Earth's radius.

For use in later chapters we should write the same equation using the more conventional names for the coordinates, θ for the co-latitude and ϕ for the longitude. Then the distance relations on a unit sphere (sphere of radius 1) are given by

$$(\Delta\ell)^2 = (\Delta\theta)^2 + \sin^2 \theta (\Delta\phi)^2. \quad (18.3)$$

If the sphere has radius r , then all distances scale in proportion to r . Since this expression gives us the squared distances on the sphere, the appropriate relation for a sphere of arbitrary radius r is

$$(\Delta\ell)^2 = r^2 (\Delta\theta)^2 + r^2 \sin^2 \theta (\Delta\phi)^2. \quad (18.4)$$

Notice that there is no term here involving Δr . All distances are measured along the sphere, at fixed r , so there are no changes in r to take into account.

Equation 18.1 on page 226 therefore gives us lots of flexibility to describe curved surfaces, but it does not have the most general form for calculating distances. It can be extended even further by putting in a term that is a product of Δx and Δy . This leads to the *most general possible* expression for the distance between nearby points on a curved surface in a given coordinate system (x, y) :

$$(\Delta\ell)^2 = A^2 (\Delta x)^2 + B^2 (\Delta y)^2 + 2C \Delta x \Delta y, \quad (18.5)$$

where A , B , and C depend on position on the surface, i.e. they are functions of the coordinates x and y .

The only difference with Equation 18.1, apart from calling the coordinates by their conventional names x and y , is the extra term with the coefficient C (which, in this context, has nothing to do with the speed of light). This coefficient corrects for the fact that the coordinate lines of x and y may not always be perpendicular to each other on a general map. This kind of thing is inevitable if the geometry is not perfectly smooth. If the map contains a mountain, for example, then there would be no way to draw the horizontal map reference lines in such a way that — if they were painted on the mountain and not just drawn on the map — they would cross the vertical map reference lines at right angles everywhere. Therefore the Pythagorean rule cannot be used in the simple form of Equation 18.1 on page 226.

It is not hard to see why this problem can be cured with the term involving C , containing a product of Δx and Δy . Remember the “cosine rule” for the length ℓ of the side of a triangle opposite to an angle θ , formed from sides of length x and y :

$$\ell^2 = x^2 + y^2 - 2xy \cos \theta.$$

When the triangle’s two sides are perpendicular, then $\theta = 90^\circ$, so that the last term is zero, leaving the usual Pythagorean theorem. But in general one needs the cosine term to compensate for the fact that the distances x and y are measured in directions that are not perpendicular to each other. We see that the cosine formula is a special case of the formula Equation 18.5 with $C = -2 \cos \theta$ and $A = B = 1$. The cosine formula can be thought of as the distance formula in flat space with straight coordinates that are skewed, so that there is an angle θ between the coordinate axes, an angle that is not necessarily a right angle.

Of course, the cosine rule is a formula that is correct only for a triangle in a flat two-dimensional plane. But every smooth curved space is locally flat (i.e. flat if we look at a small enough piece of it), so if the differences Δx and Δy are small enough, we can interpret the distance formula Equation 18.5 exactly as a version of the cosine rule. Therefore, C at any point just measures the angle θ between the directions of the coordinates at that point: $\cos \theta = -C/2AB$.

This shows that an ordinary space (not spacetime), where squared distances must be positive, cannot have a distance formula with arbitrary coefficients: C^2 must be smaller than or equal to $4AB$ (in order that $\cos \theta$ should be less than or equal to one) and A and B themselves must be positive.

Figure 18.3 illustrates a coordinate system for a two-dimensional surface that is curved. It shows that, even when the coordinates are drawn in a very regular and smooth way, they stretch and turn to follow the surface. If we choose any grid line in the diagram and move along it, we see that the distances (measured along the surface, of course) between successive intersections with other grid lines do not remain a constant length: the grid is stretched and compressed. We also see that grid lines do not always intersect at right angles. The distance formula for this surface, in this coordinate system, will have non-zero functions A , B , and C .

It is important to understand that the functions A , B , and C depend on the coordinates we have chosen, as well as on the curvature of the surface. There are many different ways I can draw coordinates on a surface, and the amount of stretching and squashing of the coordinates I need to do will depend on how I draw them, even though the surface will remain the same.

Therefore, while the functions A , B , and C contain information about the curvature of the surface, they are not uniquely determined by the curvature: they depend on the coordinate system as well.

Newtonian gravity as the curvature of time

How do we use distance measures in curved spaces to describe Newtonian gravity? We discussed curved two-dimensional surfaces because we could visualize them, but

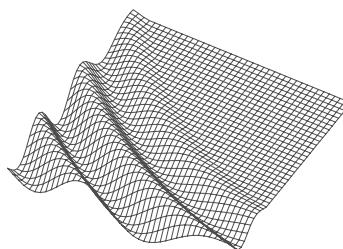


Figure 18.3. This drawing shows a simple, nearly-rectangular coordinate system drawn on a curved (wavy) two-dimensional surface. It is impossible to keep the coordinates smooth without stretching them. Generally the coordinate lines also cannot be made to intersect at right angles.

>In fact, it is possible to have complicated functions even on the flat plane, just by choosing the coordinates differently from the usual Cartesian coordinates x and y . For example, the Pythagorean theorem for small distances using polar coordinates r and θ in the plane is

$$(\Delta \ell)^2 = (\Delta r)^2 + r^2 (\Delta \theta)^2.$$

In this section: we learn how to understand the curvature of time.

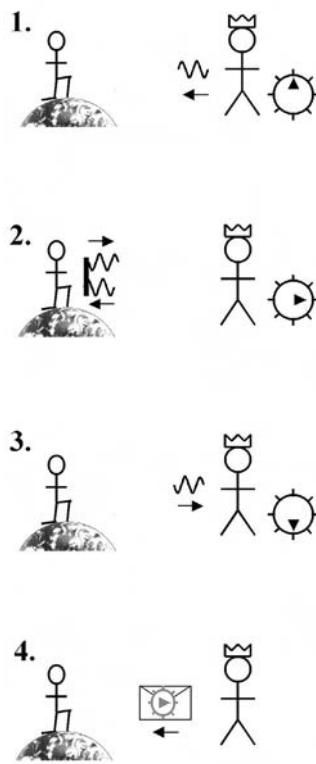


Figure 18.4. How a time coordinate could be assigned to events on Earth by a distant experimenter using his own clock. The distant experimenter (the “master”), at rest, sends a photon to his “slave” on Earth (1), who reflects it back (2). The master notes the time he receives it (3). Since light takes the same time to travel to the Earth as to return, the master assigns the average of the times of (1) and (3) to the reflection event (2) as its time-coordinate value. He tells his slave this much later, in a letter (4), but that is okay: the event has been given a unique time. If the master does this again, then the slave will notice that the elapsed coordinate time is different (longer) than the time elapsed on a clock on the Earth. Being a slave, he is in no position to object to this! Nor should he: the master determines the coordinate time to be assigned to things, even if they are not the true (proper) times. That is what a time-coordinate is.

it is harder to visualize a four-dimensional curved spacetime! The remarkable thing is that the mathematics that we have developed for measuring distances changes very little when we adapt it to describing gravity.

One big change is that in spacetime, the distance measure is the *spacetime-interval*, not the Euclidean distance, so we expect the coefficient of the time term to be negative.

The only other change is that in principle we need to use all three dimensions for space: gravity is a property of our real three-dimensional world. In the last chapter we saw how to write the spacetime-interval in just one space dimension, Equation 17.1 on page 217. To put in the other two dimensions is simple:

$$\Delta s^2 = -(c\Delta t)^2 + (\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2. \quad (18.6)$$

The extra two space dimensions y and z have the same footing as x , and for purely spatial spacetime-intervals ($\Delta t = 0$) this is the standard Pythagorean rule in three dimensions. It follows from our discussion above that a general curved spacetime is described by modifying the spacetime-interval in Equation 18.6 to add in variable coefficients and “mixed” terms. Since this could get extremely messy, we’ll only just note that it must be done to be fully general, but we won’t need to do it for our discussions! Instead, we focus here on putting a variable coefficient in front of $(\Delta t)^2$.

The key to linking the notion of curvature in time with Newtonian gravity is the gravitational redshift. We have already seen in Chapter 2 that the gravitational redshift affects the rate at which clocks run. Imagine that we establish a time coordinate in the Solar System so that the time assigned to any event is the time that is recorded by a clock far from the Sun when the event occurs. It is worth thinking a little about how this time coordinate could be set up. Let’s do it for the Earth, as a concrete example.

As we have remarked before, a clock on the surface of the Earth runs slower than one far away. How do we measure this? Let the clock on the Earth send out a radio signal each time it ticks. Then this signal will take a while to reach the distant clock, but the signals from both ticks take the same amount of time to travel, so the time between their arrivals at the distant clock will depend only on the time between their emissions: the time between ticks of the Earth-bound clock.

We define our time coordinate t even at the Earth-bound clock to be the time elapsed on the distant clock between Earth-bound ticks. This will be longer than time on the Earth-bound clock, because of the gravitational redshift. But since clocks at different altitudes on the Earth will also run at different rates, there is nothing special about the Earth-bound clock. Our global time coordinate t has the advantage that it is possible to define it anywhere.

Of course, this time is not the time that the Earth-clock ticks. Our t is just a coordinate, a way of locating events in time. It is not meant to be directly a physical measurable. The proper time, given by the spacetime-interval, is the time on the local clock.

The gravitational redshift forces us to be careful about defining time. We saw in our discussion of the GPS in Chapter 2 that we now have to take into account the differences in redshifts of different clocks in our daily timekeeping on the Earth. The gravitational redshift causes local clock time (the proper time) to be different from our time-coordinate time t , so it is exactly the factor that converts from coordinate

Investigation 18.1. The gravitational redshift tells us how time curves

In Chapter 2 we saw that the effect of a Newtonian gravitational field was to change the rate at which clocks ticked. Now, the proper time given by the spacetime-interval is the time on clocks. The coordinate time is rather arbitrary, but it is helpful to take it to be the same as the proper time of clocks that are far from the gravity of the star or black hole that we are considering. These are “our” clocks, the clocks that we as astronomers far away from the system use to measure time.

Suppose that after a given amount of coordinate time Δt , a clock at rest in the gravitational field has ticked a proper time $\Delta\tau$. The relation between these depends on the clock’s position in the gravitational field. For a clock at a distance r from a Newtonian body of mass M , a simple extension of the redshift calculation of Investigation 2.2 on page 16 shows that this relation is

$$\Delta\tau = \left(1 - \frac{GM}{c^2r}\right) \Delta t. \quad (18.8)$$

Notice that proper time and coordinate time are equal when we are far away from the star or black hole ($r \rightarrow \infty$), which is how we defined the time coordinate t . This equation is only valid if the Newtonian field is weak, i.e. if $GM/c^2r \ll 1$.

Now, along the world line of a clock that is at rest, the spatial coordinates don’t change, so if we use the spacetime-interval to calculate the proper time, we can set $\Delta x = \Delta y = \Delta z = 0$. The negative of the spacetime-interval Δs^2 is the square of the proper time, $\Delta\tau^2$, times c^2 , so we are led to

$$\Delta s^2 = -\left(1 - \frac{GM}{c^2r}\right)^2 (c\Delta t)^2. \quad (18.9)$$

Exercise 18.1.1: Redshift near the Sun

Derive Equation 18.8, starting from Investigation 2.2 on page 16. Calculate the redshift experienced by a photon with a wavelength of $0.5 \mu\text{m}$ as it travels from the surface of the Sun to a very distant observer. Calculate the redshift of the same photon if it is observed by a space observatory in the Earth’s orbit but far from the Earth. Finally, calculate the redshift if the same photon is observed by an astronomer on the surface of the Earth.

time to proper time. The gravitational redshift is therefore precisely the “squash-ing” factor that we are looking for in the spacetime-interval formula for the Solar System. The details of how to put the redshift factor into the spacetime-interval are worked out in Investigation 18.1. The result is a spacetime-interval where only the time-coefficient is variable:

$$\Delta s^2 = -\left(1 - \frac{2GM}{c^2r}\right)(c\Delta t)^2 + (\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2, \quad (18.7)$$

where M is the mass of the Newtonian star (the Sun could be replaced in this argument by any star) whose gravity we represent by this curved spacetime, and where $r^2 = x^2 + y^2 + z^2$ is the distance from the star to the point (x, y, z) in space where the clock is.

Do the planets follow the geodesics of this time-curvature?

The test of whether our expression Equation 18.7 for the spacetime-interval represents the real world is whether this spacetime has geodesics that are the Newtonian trajectories of particles orbiting the mass M . This may again sound like a hard thing to show, but it is not. In fact, we have essentially done all the work we need to show this. We just have to assemble the various components of the argument.

The key is to realize that the locally flat coordinates in a spacetime are the coordinates of observers who fall freely with the acceleration of gravity. These observers can, by the equivalence principle, expect to do local experiments and have them come out exactly as in special relativity, as if there were no gravity. Now, one of the ex-

The term that is squared can be simplified by expanding the square:

$$\left(1 - \frac{GM}{c^2r}\right)^2 = 1 - 2\frac{GM}{c^2r} + \left(\frac{GM}{c^2r}\right)^2.$$

The last term on the right-hand side is very small, since we are assuming the gravitational field is weak. For example, on the surface of the Earth we have $GM/c^2r \approx 10^{-8}$, so the last term is about 10^{-16} as large as the second term. We will neglect it now. We must not throw away the second term, however, since that is where all the deviations from special relativity occur! We get

$$\Delta s^2 = -\left(1 - \frac{2GM}{c^2r}\right)(c\Delta t)^2. \quad (18.10)$$

The time part of the spacetime-interval is therefore determined by the gravitational redshift effect. For the spatial coefficients, we make the simplest assumption: keep them the same as in special relativity. This gives the spacetime-interval Equation 18.7. We will see that this works perfectly for geodesics that represent particles going at non-relativistic speeds.

By the way, don’t think that Equation 18.9 is more accurate than Equation 18.10, just because we have dropped the squared term to get the result. In fact, Equation 18.9 is itself not fully correct when gravitational fields are strong, so it can have errors of the same general size as the term we have neglected. Ironically, it turns out that Equation 18.10 is actually the form that is correct for strong fields as well as weak ones.

▷ Note how different a time coordinate is from a time measurement. To *measure* the time passing on Earth, the clock must be on the Earth too. To assign an arbitrary *time-coordinate* one can use any clock, and it is particularly convenient to use one so far away that the Earth’s gravity does not slow it down.

In this section: we show that the motion of any object acted on by a Newtonian gravitational force can be fully described instead as a free motion along a geodesic of a geometry with curved time.

periments they can do is to watch another nearby freely-falling particle. Since there is no gravity in their local (freely-falling) spacetime coordinate system, this particle will move on a straight line through their coordinates.

But this is the definition of a geodesic: a geodesic is a straight line in a locally flat coordinate system. Therefore, the geodesics of a spacetime in which the locally flat coordinates are those of experimenters falling freely in a gravitational field are the trajectories of freely-falling particles in the same gravitational field.

▷ Fundamentally, we have turned our old derivation of the gravitational redshift completely around. In Chapter 2 we derived the redshift from the Newtonian gravitational force. In the present chapter, we have derived the Newtonian "force" (really, the equivalent spacetime geometry) from the gravitational redshift. From Einstein's point of view, the redshift is the more fundamental of the two, since it directly measures the geometry of spacetime. The motion of particles follows almost incidentally from that geometry.

This proves that Equation 18.7 on the preceding page describes a spacetime geometry in which particles that follow geodesics will move on exactly the same trajectories as particles would do in a flat spacetime with the Newtonian gravitational force acting. We have therefore found a curved-spacetime picture of Newtonian gravity. The curvature here is *only* in the time-direction. Curvature in time is nothing more than the gravitational redshift: time advances at different rates in different places, so time is curved. We have found that the gravitational redshift fully determines the trajectories of particles in the gravitational field.

We have arrived at this goal with a minimum of calculation. We did not have to do any calculus or solve any differential equations. Yet we now know what it means physically when we say that time is curved: it means that the rate at which clocks run changes from place to place, even when the clocks are at rest with respect to one another. The curvature of time is in the gravitational redshift, and the gravitational redshift is enough to insure that freely-falling bodies follow their Newtonian trajectories.

All of Newtonian gravitation is simply the curvature of time.

How to define the conserved energy of a particle

The frequency of photons changes because of the redshift, so their energy also changes. Nevertheless, it is possible to define a conserved energy in relativity, just as it was in Newtonian gravity. This should not be surprising, especially considering that energy depends not only on the particle but also on who is measuring it.

Let us identify the experimenters who measure the redshifted energy of a photon. They are local experimenters who are at rest with respect to the star. They each perform a local experiment to measure the frequency of the light, and they find that it decreases as the photon climbs away from the Earth. If they had been observing a freely-falling particle they would have found a similar result: the speed, and hence the kinetic energy, of a particle gets lower and lower as it gets further and further from the star.

An observer very far away, so far that GM/rc^2 is too small a correction to measure, is in a special position: this is where we are when we observe almost all astronomical systems outside the Solar System. In an ideal case, where we consider only the gravity of a single star, this distant observer lives for all practical purposes in the spacetime of special relativity. In special relativity, the energy of a particle or photon is constant as it moves. Although this is not true of the photon that moves near the star, it becomes true when that photon moves far from the star: as it leaves the gravitational influence of the star, its energy (frequency) becomes constant, regardless of where it is going.

This energy, as measured by a distant observer, is called the *conserved energy* of the photon. It is conserved in what might seem to the reader to be a trivial sense, in the sense that it is a number that is *defined* to be a constant equal to the energy when the photon is far away. If we associate this number with the photon even

▷ Actually, although we are far from other astronomical bodies, we do sit deep in the Earth's gravitational field. Astronomers agree to use a universal time coordinate that matches proper time on the Earth, not in interstellar space.

▷ The conservation of the energy we have defined is not actually trivial, as explained below.

when it is near the star, this does not mean that it is measurable near the star. It is, rather, a property of the photon (or particle) and the overall spacetime, which corresponds to the result of doing a direct measurement on the photon when it is far away.

Now, this might seem an odd thing to define, but let us see how it works out. The way to define this energy is to apply the redshift effect, to multiply the energy as measured by any local experimenter near the star by the right factor to get the energy as seen far away. Since the redshift is just a change in clock rates, and since we now have a spacetime-interval in Equation 18.7 on page 231 that describes clock rates, the definition of the conserved energy is straightforward:

$$\text{conserved total energy} = \text{local measured energy} \times \frac{\sqrt{|\Delta s^2|}}{c\Delta t}.$$

Let us write E_{local} for the locally measured energy and $E_{\text{conserved}}$ for the conserved total energy. Then using the spacetime-interval we get

$$E_{\text{conserved}} = \sqrt{1 - \frac{2GM}{c^2r}} E_{\text{local}}. \quad (18.11)$$

Since we are assuming that GM/c^2r is a very small number, we can use the binomial approximation in Equation 5.2 on page 43 to replace the square-root in this equation by $1 - GM/c^2r$. It is not hard for photons to put in the energy $E = h\nu$ and get back the gravitational redshift formula that we started with.

More interesting is what this implies for particles that do not travel at the speed of light. Their local energy is just the energy that a special-relativistic experimenter would measure, so if they have speed v at some location, then their local energy is given, as we have seen in Chapter 15, by their rest-mass mc^2 times the relativistic gamma-factor $\gamma = (1 - v^2/c^2)^{-1/2}$. If the velocity is slow, then using the binomial approximation again gives $\gamma = 1 + v^2/2c^2$, and the total energy is

$$E_{\text{local}} = mc^2 + \frac{1}{2}mv^2.$$

Here we see clearly that the local energy is just the kinetic energy of motion, plus of course the rest-mass energy. The total energy is this times $1 - GM/c^2r$. This multiplication involves two expressions, each containing two terms. Therefore the result has four terms. However, one of them is the product of the two very small quantities, and is of the same size as the terms we dropped in using the binomial approximation. So the final result is, to the accuracy we have been working, just

$$E_{\text{conserved}} = mc^2 + \frac{1}{2}mv^2 - \frac{GMm}{r}. \quad (18.12)$$

This is, apart from the constant rest-mass term, exactly the form of the conserved energy that we had in Chapter 6 by adding up Equations 6.8 and 6.9 on page 54.

We have learned something deep here. The constant energy of an object orbiting the Sun is the locally measured energy it has when its trajectory takes it far from the Sun. The relativistic total energy is the generalization of this concept to relativity. However, some orbits in Newtonian systems cannot get far away; they are said to be bound to the central star. These include the normal circular orbits of planets. They have negative Newtonian energy, which means their total relativistic energy is less than their rest-mass energy. This proves that they cannot escape to distant regions on these trajectories: no object could have a locally measured energy less

than its rest-mass. But the total conserved energy can nevertheless be defined by asking what extra energy would be required to get the planet far away with zero speed, so that its locally measured energy far away is its rest-mass energy. Then the actual energy of the object on the bound orbit is its rest-mass energy less this escape energy. In this way, all conserved energies are measured with respect to a distant experimenter.

In relativity, when speeds are not slow and gravitational fields are not weak, the full energy will be much more complicated than Equation 18.12 on the previous page. However, as long as the geometry of spacetime does not change with time, it is possible to define the conserved energy.

►If the geometry were to change with time, then the energy of the particle when it arrives at the distant observer will depend on just what wiggles and changes in the geometry it has encountered on the way. The energy will not be a constant since a different particle starting out the same way could finish with a different energy. We saw in our discussion of the gravitational slingshot in Chapter 6 that energy can only be conserved if the forces are time-independent. This applies to the geometry of spacetime as well.

The conserved total energy of a particle is especially important because it is not only constant along the trajectory of the particle, provided it falls freely (follows a geodesic), but it is also conserved for a collection of particles. This follows from the local conservation of energy. Consider two particles that collide. Before their collision they each have a conserved energy and it is constant, so their total conserved energy is constant. When they collide, their local energies change, but only in such a way that the total local energy is constant. The local and conserved total energies differ from one another only by terms that depend on position, like GMm/c^2r . Since the collision takes place at a particular position, these terms are the same for the two particles, so that their total conserved energies after the collision add up to the same value as before. This argument can clearly be generalized to many particles or to a continuous body.

If the geometry is time-independent, then the total energy of a material system, as measured by an experimenter very far away, is conserved, independent of time.

The deflection of light: space has to be curved, too

In this section: we calculate, with simple algebra, the correction to the Newtonian deflection of light that is caused by the curvature of space.

The measured deflection tells us how space must be curved near the Sun.

We made remarkable progress in fashioning gravity as geometry by our discussion of the curvature of time. Had Einstein wanted just to describe Newtonian gravity, he could have stopped there. But Einstein already knew from special relativity that there is no unique way to distinguish time from space. If time is curved, as measured by one experimenter, then space will be curved to another. What physical effects follow from the spatial curvature?

In Equation 18.7 on page 231 the spatial coordinates obey the usual three-dimensional Pythagorean theorem everywhere, so they are coordinates of flat Euclidean space. To introduce spatial curvature, we must put other coefficients in front of $(\Delta x)^2$ and its relatives.

Why did we not go to this complication already in Equation 18.7? Did we make an over-simplification? After all, maybe the spatial coefficients are similar to the one for time. Would they not affect the geodesics so that they do not reproduce the Newtonian orbits? The answer, profoundly, is no.

We can neglect the coefficients in the spatial part of the spacetime-interval because they actually have little effect on the motion of a planet.

The reason for this is the slow speed v of a Newtonian gravitational orbit. In a given time ΔT (as measured in meters), the distance traveled by the orbiting body is $(v/c)\Delta T$, which is much less than ΔT . Therefore, the time coefficient in the spacetime-interval dominates the value of the interval for planetary motion: the other terms in the spacetime-interval are small, and therefore the effects produced by their slight deviations from one can have only a slight effect on the Newtonian motion.

When we come to look at the motion of photons, however, the rest of the metric does matter, because the contributions to the spacetime-interval of the terms involving Δx are comparable to those from ΔT .

The motion of photons is sensitive to the curvature of space, and measuring what photons do can tell us what corrections we have to put into the spatial part of the spacetime-interval.

Since the spacetime-interval in Equation 18.7 on page 231 produces exactly the same geodesics as a Newtonian gravitational field, the geodesic of a photon that passes a distance d from the Sun *in this geometry* will be deflected by the Newtonian value of $2GM/c^2d$, as given in Equation 4.13 on page 38. But, as we saw in Chapter 4, observations of the 1919 eclipse showed that the deflection was actually twice this:

$$\text{angle of deflection of light by the Sun} = \frac{4GM_{\odot}}{c^2d}. \quad (18.13)$$

We must, therefore, find appropriate corrections to the spatial part of the spacetime-interval. The corrections must of course involve the quantity GM/c^2d , since it is the only physical quantity available.

Again, although the problem of determining what this correction should be sounds difficult at first, there is a plausible argument that will allow us to guess the result. Let us put an arbitrary correction term into Equation 18.7 on page 231 of the form we expect. This amounts to considering a metric of the form

$$\Delta s^2 = -\left(1 - \frac{2GM}{c^2r}\right)(c\Delta t)^2 + \left(1 + \gamma \frac{2GM}{c^2r}\right) [(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2], \quad (18.14)$$

where γ is a constant whose value we have to determine.

This metric is already much simpler than a general one with spatial corrections would be: I have taken the coefficients of all three spatial coordinate changes to be the same, and I have not allowed any “mixed” terms. Since we are looking for a small correction to flat space, we take this function to be $1 + 2\gamma GM/c^2r$.

To determine γ , remember that, for a photon, the time difference (in distance units) $\Delta T = c\Delta t$ between two events will roughly equal the spatial distance. Therefore, the spatial corrections to the spacetime-interval will have the same “weight” as the corrections to the time part: when we add up distances to get the spacetime-interval, both corrections are multiplied by the same $(\Delta T)^2$ and then added in. Moreover, since the sign of the time part of the spacetime-interval is negative, the spatial correction will *increase* any physical effect caused by the time part if it has the *opposite* sign to the time correction. I have written the corrections in Equation 18.14 with opposite signs already, so we can conclude from this that, if the time part alone creates a certain deflection of light, then the time and space corrections together will make a deflection that is $1 + \gamma$ times as large. A full calculation of the trajectory of a photon confirms this conclusion.

We have seen that the observations indicate that the deflection of light is $4GM/c^2$. This is twice the Newtonian value, i.e. twice the value that one would get from only the time correction in the spacetime-interval. From this we conclude that $1 + \gamma = 2$, or $\gamma = 1$.

The full spacetime-interval for the spacetime curvature created by a star like our Sun is, therefore,

$$\Delta s^2 = -\left(1 - \frac{2GM}{c^2r}\right)(c\Delta t)^2 + \left(1 + \frac{2GM}{c^2r}\right) [(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2]. \quad (18.15)$$

►This use of the symbol γ has nothing to do with our other use of γ in the time dilation or Lorentz–Fitzgerald contraction formulas. I have used γ here because it has become the conventional symbol physicists use for this constant when they discuss measurements of the spatial curvature in the Solar System.

►Excluding mixed terms in the interval is justified in this case by the fact that the spatial geometry must be spherically symmetric: the Sun is round, and I should be able to rotate my x - y - z coordinate system in an arbitrary way about the Sun’s position without changing the form of the spacetime-interval. This means that the spatial part of the spacetime-interval should look just like flat space except for being multiplied by an arbitrary function of the distance r from the Sun.

►Our argument used only photons that passed by the Sun, so we cannot use it to deduce the geometry inside. That's good: we don't have to ask, at least not yet, what happens when r is small enough to change the sign of the coefficient of $(\Delta t)^2$. But we *will* ask this question in Chapter 21.

In this section: our estimate of the curvature of space sets a challenge to any geometrical theory of gravity that wants to replace Newton's. Different theories can be expected to predict different curvatures. Einstein found that his theory predicted the correct curvature automatically.

Apart from the approximation that we should be far enough from the star that the effects of gravity are weak enough to be represented by Newtonian gravity, this spacetime-interval is a full representation of the geometry outside the Sun. Its curvature determines the motion of any particle, from a slowly moving planet to a speedy photon. That we were able to derive it from the observed deflection of light past the Sun, again with only a minimum of mathematics, is testimony to the fundamental simplicity of Einstein's ideas on gravity.

Space curvature is a critical test of general relativity

We have calculated both the curvature of time and the curvature of space in the Solar System, but we should not go further without noticing that there was a big difference between the premises we used to get them. In this difference lies the explanation of why the measurement of the deflection of light by the Sun was so important for the acceptance of general relativity as the correct theory of gravity.

We derived the curvature of time from the gravitational redshift and the equivalence principle: it followed directly from Einstein's basic picture of gravity, and especially from the importance he gave to the equivalence principle. It is therefore the case that *any* relativistic theory of gravity that embraces the equivalence principle in this geometrical way will have the same curvature of time that we derived. The gravitational redshift is the key that opens the door to the geometrical description of gravity. But it leads to no new physical effects all by itself.

It is important to realize that Einstein's theory consists of more than just his picture of gravity as spacetime curvature. The key part of his theory is what are called the **Einstein field equations**, which tell physicists how to calculate the corrections to the spacetime-interval when they know what systems create gravity in any particular situation. Two different theories of gravity could both adopt the geometrical point of view, give the same curvature of time, and yet make very different predictions about the curvature of space.

So far, we have deduced the curvature of space, not from a general principle, but from the *observations* of the deflection of light by the Sun. If the observations had given a deflection three times the size of the Newtonian one, then we would have set γ to 2 and been just as happy with it. In fact, it is the job of the theory of gravity to *predict* this deflection, but this can be done only when we use the theory to calculate the curvature that the Sun should produce. This is what the Einstein equations are designed to do, and we will show in the next chapter that they do indeed predict $\gamma = 1$.

Other scientists, before and after Einstein, have suggested other equations for gravity, usually more complicated, and they usually produce different kinds of spatial curvature.

The spatial curvature, as measured by the deflection of light, is a key *test* of the particular theory of gravity, not just of the general geometrical framework in which we describe gravity.

The remarkable fact about general relativity is that – as we shall see – it predicts that the spatial curvature produced by the Sun will be exactly as given in Equation 18.15. Einstein did not know about the size of the effect when he devised general relativity: it was only measured three years later. The measurement of the deflection, showing that the spatial curvature was exactly as his theory predicted, coming on top of his earlier explanation of the precession of the orbit of Mercury (see the next section), was the event that propelled Einstein into superstardom, making his name a household word for “genius”.

How Einstein knew he was right: Mercury's orbital precession

Einstein was convinced of the correctness of his theory even before the deflection of light was measured. The reason is that the theory had already explained the tiny anomaly in Mercury's orbit: the extra precession of its perihelion.

We have seen that photons respond more to the curvature of the spatial part of spacetime than do planets, because they move faster: the contribution of Δx to the total spacetime-interval for a photon is similar in size to that of ΔT . Planets, on the other hand, have much smaller values of Δx for a given time spacetime-interval ΔT , so the effect of spatial curvature on their orbits is smaller. But it won't be exactly zero, and the effect it has will be something new, not contained in Newtonian gravity. Can we estimate what the spatial terms might do to an orbiting planet?

It seems reasonable to expect that the effect of spatial curvature on a planetary orbit would be similar to that on a photon's trajectory, at least qualitatively. For the photon, spatial curvature increased the deflection, drawing it further away from its original direction and toward the Sun.

So let us consider what might happen to a particle that is in an elliptical Newtonian orbit around the Sun. Its orbit should, like the photon's, turn a little bit more than the Newtonian orbit would. This would mean that the orbit would no longer be exactly a closed ellipse: in the time it takes the planet to go from one perihelion to the next, the orbit will turn by a little more than 2π . This is exactly what is observed for Mercury: the perihelion *advances* by a small amount each orbit, as we saw in Chapter 5.

The extra angle per orbit should be similar to the angle by which the photon is deflected by the spatial curvature, $2GM/c^2r$, essentially because this is the only physical number that is available in this problem. But it need not be exactly this value. For one thing, the photon experiences the deflection as it passes by the Sun on one side, while a planet goes all the way around. Moreover, there may be other effects besides spatial curvature that could induce a perihelion precession. For example, we derived the form of the spacetime-interval in Equation 18.15 from the Newtonian motion of the planets. Einstein's theory might add further, smaller correction terms. These could have the form, for example, of adding $(GM/c^2r)^2$ times some extra coefficient into the time coefficient in the spacetime interval.

Einstein's theory does indeed predict such extra terms (the "extra coefficient" of the previous sentence turns out to be two), just as it predicts the spatial curvature. We cannot calculate them here from what we know of the theory so far, but when all such terms are taken into account, Einstein's theory predicts that the perihelion position of a nearly circular orbit will advance by an angle of

$$\delta\phi_{\text{peri}} = \frac{6\pi GM}{c^2 r} \quad (18.16)$$

per orbit, a factor of $3\pi/2$ times the total angle by which light is deflected. If we evaluate this angle using numbers appropriate for Mercury's orbit (setting M to the mass of the Sun and r to the radius of Mercury's orbit, which we can obtain from Table 4.2 on page 28), we get an advance for each orbit of 4.8×10^{-7} radians (just under 0.1 arcseconds). Astronomers don't have the accuracy to detect this advance on each orbit, but fortunately they don't need to. Since the same advance occurs for each orbit, the successive advances just accumulate. In 100 years, Mercury makes about 415 orbits, so the accumulated perihelion advance is an easily measurable 41 seconds of arc per century.

This is almost exactly the extra perihelion advance of 43 arcseconds per century that had been known but unexplained for decades before Einstein arrived at general

In this section: even before the curvature of space was measured using light deflection, Einstein had shown that the spatial curvature predicted by his theory explained the anomalous precession of the orbit of Mercury.

relativity. (See the discussion in Chapter 5.) If we had used a more accurate formula that included the slight effect of the eccentricity of Mercury's orbit on the prediction of general relativity in Equation 18.16 on the preceding page, then agreement would be within the observational errors, which are less than one arcsecond per century. This is a very satisfactory agreement.

Einstein was well aware of the problem of the unexplained perihelion advance while he was working on general relativity, and when his development of the theory had gone far enough for him to calculate the effect on Mercury's orbit, he immediately found the value in Equation 18.16.

By this time, his theory had no unknown constants in it, so there was nothing left to adjust to fit the observations: either the theory predicted the right amount of precession or it didn't. It did.

The prediction of the correct result for the precession of the perihelion of Mercury, coming at the end of a long and painstaking calculation, gave Einstein palpitations of the heart. In his own words: "For a few days I was beside myself with joyous excitement." He knew then that his theory was the right one.

The perihelion advance and the extra non-Newtonian part of the deflection of light are only two of a number of new effects in the Solar System that Einstein's theory predicts. These are called **post-Newtonian** effects: small corrections to Newton's predictions. A number of post-Newtonian predictions of Einstein have been verified to accuracies of better than 1%. It is important to stress that Solar System tests are weak-field tests, so they leave open the possibility that the correct theory of gravity is one that differs from Einstein's only for very strong fields, like those near a black hole. Moreover, we expect that the theory will fail when quantum effects are important, since the theory does not take quantum gravity into account at all. We will return to this theme in later chapters.

Weak gravity, strong gravity

In this section: we look ahead at the task of the next chapter, to show how the curvature of spacetime is generated by the bodies in it.

The discussion in this chapter and the next, though based on weak gravitational fields, will explain a great many things about strong-field situations in general relativity. These situations, involving compact objects or cosmology, are the places where general relativity is most needed, and where it predicts and explains phenomena that even Einstein never dreamed of when he wrote down his equations. We will begin our study of these in Chapter 20. But first we need to complete our journey through relativity by learning how to work out the gravitational fields that different objects will create. We need to learn about Einstein's field equations.

Einstein's recipe: fashioning the geometry of gravity

We are now ready to go to the heart of general relativity, to learn how matter *generates* gravity. This subject is usually left out of discussions of general relativity below the level of an advanced university course. The reason is mathematics, not physics: Einstein formulated his field equations, his gravity-generating equations, using the language of differential geometry. This is the mathematical discipline that deals with curvature, and it is far from elementary. The physical ideas that Einstein expressed in this mathematical language are simply too important, however, to pass over. In this chapter we whittle down the mathematics to a form that is as close as possible to the algebra we used in our earlier chapters on Newton's gravity. This allows us to share in Einstein's thinking, to see what general relativity really predicts about the world we live in.

We are stepping here into a realm that is amazingly rich with new ideas. We will see how Einstein introduced a modest change in the way that the Newtonian part of the gravitational field is generated, a change that led, step-by-step, to modern concepts like cosmological **inflation** and the accelerating Universe. We will see how stars curve the space they live in, by just enough to explain the observed deflection of light as it passes the Sun. We shall see why gravity in special relativity implies that moving bodies create a new kind of gravitational effect, called **gravitomagnetism**. This effect is about to be tested in an experiment in orbit, a test that will (coming full circle) also probe the foundations of the theory of cosmological inflation. The phenomena we meet here form the basis of all the remaining chapters: why stars collapse to black holes, why rotating black holes can power quasars, how binary systems radiate gravitational waves, how the expansion of the Universe can be accelerating – all these and more are grounded in an understanding of how gravity is created.

Not that the mathematics of differential geometry is unimportant: far from it. Einstein's equations are quite beautiful when expressed this way. Not only are they compact (as we will see below) but they have a very deep symmetry: they have exactly the same form in any coordinate system, so they give equal status to any observer/experimenter, inertial or not. This extension of the principle of relativity is called the **principle of general covariance**. Where the principle of relativity placed all *experimenters* on the same footing, the principle of covariance says that the field equations must be able to be used in any *coordinate system*, no matter how peculiar. This is rightly regarded as a beautiful and powerful aspect of Einstein's theory.

To make the mathematics simpler, we have to set aside general covariance, at least temporarily. We shall analyze how matter creates gravity from the point of view of a particular observer, an observer who is essentially at rest with respect to the sources of gravity. We shall follow the pattern of the previous chapter, where we separated the curvature of time from the curvature of space. Here we will look

In this chapter: we study the equations that show how matter generates gravity in general relativity. We identify four properties of matter and gravity that act as sources of gravity, and we show how these different sources produce different gravitational effects. Using only a little algebra, we compute the curvature of space and get the observed deflection of light as it passes the Sun. We show how special relativity and the curvature of time lead to something called the dragging of inertial frames. We examine the special properties of the cosmological constant as a source of gravity.

>The image beneath the text on this page is from a numerical simulation of the merger of two black holes. It shows a measure of the curvature of space that carries information about gravitational waves. In the center can be seen two small blobs surrounded by a larger one: these are the two original black holes after they have merged into a single one. The rest of the picture shows gravitational waves moving out. Image by W. Benger, Zuse Institute Berlin (ZIB), from a simulation performed by scientists at the Max Planck Institute for Gravitational Physics (Albert Einstein Institute, AEI), Washington University (WASHU), and the National Center for Supercomputer Applications at the University of Illinois (NCSA).

►Our picture of general relativity builds on insights gained by generations of physicists – successors to Einstein – who took apart the complex field equations and painstakingly won the physical insights and developed the physical intuition that are necessary for applying general relativity to the real world of astronomy.

In this section: we list the four distinct sources of gravity in general relativity: active gravitational mass, active curvature mass, momentum, and gravity itself.

►When something physical can help create itself, as in item 4 in the list, physicists say that it is **non-linear**. This refers here to the relation between the gravitational field and its source. In Newton's gravity, if the density of mass inside one star is twice that of another, while their sizes are the same, then the gravitational field of the first will be twice as strong as the second. The field will be proportional to the density; a graph of the field strength against the density of the star will be a straight line, which mathematicians describe as a **linear** relationship.

That does not happen in general relativity. The larger density will create a larger field, of course, but then this larger field will create even more energy, and the result will be that the field of the denser star is not just twice as strong as that of the other. The graph of the field strength against the density will not be straight line: it will be non-linear. It is even possible in general relativity to have a pure gravitational field, with no matter at all, acting as its own source!

separately at how matter generates the curvatures of time and of space. We shall learn how to compute the geometry of gravity when gravity is weak, and we will build on this insight to understand how strong gravity works.

Of course, if the insights we obtain by simplifying the mathematics in this way are valid, then we should expect that there must be a sense in which they are independent of observers. After all, if one observer predicts that a star should collapse and it does, then all other observers must have been able to make that prediction as well, based on their own measurements. So at the end of our discussion we will look again at the principle of covariance that guided Einstein, use it to draw further insights into the theory, and return to the very real beauty of the theory that has been called the greatest creation a single human mind has ever achieved.

Einstein's kitchen: the ingredients

You can think of this chapter as an excursion into the kitchen of general relativity, where we will see how the gravitational fields of general relativity are made. Creations from this kitchen, such as black holes and cosmology, are displayed in every popular book on relativity, but you don't always get to go into the back room and see how they are cooked up. In later sections we will study Einstein's recipe, how he begins with the fundamental ingredients and arrives at the finished product. The recipe is called Einstein's equations. We start here with the foundation of any good recipe: the ingredients.

The ingredients are what physicists call the *sources of gravity*: things, like mass, that are responsible for gravity. The richness of the predictions of general relativity comes directly from the rich variety of sources of gravity that Einstein uses.

Newton believed that only the masses of objects create gravity. This was a powerful principle, which worked well for two and a half centuries. But, with only one ingredient, the result was inevitably a little monotonous. By contrast, Einstein's gravity involves at least four main ingredients, four distinct kinds of sources. We list them here and explain them in subsequent sections:

1. The **active gravitational mass** plays the role in Einstein's gravity that the ordinary mass plays in Newton's: it produces the main gravitational effect, namely the curvature of time.
2. The **active curvature mass** generates the curvature of three-dimensional space, which is totally absent from Newtonian gravity.
3. The ordinary *momentum* of matter generates what physicists call *gravitomagnetism*, the part of gravity that acts on masses in a way that resembles the way magnetism affects charged particles.
4. *Gravity itself creates gravity*. This is inevitable, since energy has mass in relativity, and even in Newtonian gravity there is an energy associated with the gravitational field, which we called the gravitational potential energy. Thus, gravitational fields have energy and this feeds back into the gravitational field.

The fourth source is fundamental to Einstein's picture of gravity, but it makes the equations hard to solve. You think you have a solution, but you have to change it to take into account the way your solution acts as part of its source. Normally this kind of problem is solved using computers. The huge variety of possible gravitational fields is now being explored numerically by physicists who use the most powerful available supercomputers. Because of the non-linearity of general relativity, we will only explore in this chapter how Einstein's equations create relatively weak gravitational fields. For weak gravity, the contribution of the field itself to

making more field can be ignored compared to other sources, and we are back to a linear problem.

These four sources are enough for us to gain deep insights into the way gravity works in general relativity, without using sophisticated mathematics. What we lose by ignoring the mathematics is the principle of general covariance, the equivalence of all observers. This is not just an abstract idea. It simplifies many of the concepts. It enabled Einstein actually to postulate only *one* non-gravitational source for the gravitational field, which we will return to later. The symmetry of the principle of covariance provides relationships among the first three of our sources, implying for example that if the energy density is a source, then the momentum density must also be a source. These three sources of gravity are not independent, not chosen arbitrarily. Einstein was led to them because they are different aspects of one mathematical object. So the price we pay for ignoring general covariance is that we don't see this deeper layer of unity in the theory.

In fact, the principle of general covariance insures that we are safe even if we ignore it! Since any observer, any coordinate system, is a valid system for describing gravity, the point of view of the observer we will use – one who is at rest with respect to the systems we study – is just as good as that of any other. And since any gravitational field is weak in the neighborhood of a freely-falling observer (the equivalence principle), even our assumption of weak fields is not as drastic as it might first have seemed. In fact, our four sources and our prescriptions for generating gravity from them will be excellent guides even to strong-field gravity, like black holes. Even when we study cosmology, where the gravitational fields are strong enough to control the whole Universe, we will be able to explain the expansion and acceleration of the Universe entirely in terms of weak-field gravity.

So let's get going. Welcome to Einstein's kitchen!

Einstein's kitchen: the active gravitational mass comes first

In Newton's theory, gravity is created by mass. So in general relativity, we should expect that the main source of the curvature of time will be mass. Indeed, in the last chapter we used the mass of the Sun to obtain the spacetime geometry of the Solar System. We were able to compute the effect of gravity on a relativistic particle, namely a photon as it passes the Sun, but the source of gravity – the Sun – was non-relativistic. What would gravity be like if its source were more relativistic?

Even the Sun can be made relativistic: just view the Solar System from a rocket ship moving past at $0.5c$. Would such an observer still be able to use Newton's prescription to calculate the gravitational redshift due to the mass of the Sun and to predict the motion of, say, the Earth around the Sun?

Immediately we see a problem: "mass" has no unique meaning in relativity. Rest mass is an invariant, but it is not really a suitable source of gravity. It is not even particularly well-defined for a composite object like the Sun; do we mean the total mass of the Sun as measured when it is at rest, or the sum of the rest-masses of all the particles, each as measured by an experimenter at rest with respect to it? These are different, because (as we saw in Chapter 16) the total mass of the Sun at rest includes the energy of the photons and the kinetic energies of all the particles, along with their rest-masses. So rest-mass is not a suitable source.

A better relativistic generalization of Newton's mass is the total mass-energy; this is at least conserved during nuclear reactions. But the mass-energy of the Sun is not an invariant in relativity, so the observer in the rocket will calculate a very different value for it. We might therefore expect that other properties of the Sun, which may also depend on its velocity, might be sources.

What can these properties be? When moving, the Sun will have a momentum,

In this section: the active gravitational mass generates the curvature of time, which is the most important part of the geometry of gravity. Its density is defined as the density of ordinary mass-energy, plus three times the average pressure divided by c^2 .

▷ Rest mass has another problem. What would happen to a gravitational field created by rest-mass when rest-mass is turned into energy by nuclear reactions? Would gravity disappear? This seems unreasonable. Rest mass is a dead end.

EINSTEIN SIMPLIFIED

Figure 19.1. Although we simplify Einstein's equations in this chapter by presenting them only for weak gravitational fields and only for a particular kind of observer, we retain the essential way that they link curvature to the properties of the matter that creates it. The simple form helps us to understand how relativistic gravity works even when the field is strong. Simplify, but don't go beyond the point of recognition! Cartoon copyright S Harris, reprinted with permission.



► We shall keep momentum on the ingredient list, however, and use it later as a source for other parts of the gravitational field.

so this would certainly be a candidate. On the microscopic scale, the random momentum of the gas particles inside the Sun gives it pressure, and in fact the overall motion of the Sun past the rocket observer gives it something that physicists call **ram pressure** (which we will define below), so we might expect that both momentum and pressure could contribute to the gravitational field. Now, momentum has a direction, so when we talk about a source for the curvature of time, which has no spatial direction, momentum is not a candidate. But pressure is. And in Einstein's theory, pressure does indeed play a key role in creating the curvature of time.

If we write Einstein's equations down from the point of view of an observer who is at rest with respect to the Sun (or another body that is creating gravity) then we find that the source of the curvature of time is the active gravitational mass:

density of active gravitational mass

$$= \text{density of total mass} + 3 \times \text{average pressure}/c^2. \quad (19.1)$$

Here the term "total mass" means all energies added together, and converted to their mass equivalent; and "average pressure" means the pressure averaged over the three directions in space. In a gas, pressure is the same in all three directions, so the average pressure is just the ordinary fluid pressure. We will later meet other situations where the averaging gives a different result because the pressure is not isotropic.

In Newtonian situations the contribution of the pressure is negligible, essentially because, as we saw in Investigation 7.2 on page 78, pressures of gases are typically of the same size as their densities times the square of the random velocities of gas molecules. When one divides this by c^2 as in the above equation, the term is much less than the density of mass itself. So the extra gravity produced by pressure was not noticed before Einstein.

Einstein's kitchen: the recipe for curving time

In this section: the active gravitational mass curves time in our approximation just as in Newtonian gravity.

Knowing that the source of time curvature is the active gravitational mass already gives us some insight into relativistic gravity, but in many cases one wants to be able to compute the curvature of time explicitly. Here we shall spell out the rule for computing the coefficient of $(\Delta T)^2$ in the spacetime-interval, when gravity is not

too strong, so we don't have to worry about the gravitational field as an additional source.

The rule is similar to, but actually simpler than, the rule we used in Investigation 4.3 on page 35 to compute the Newtonian gravitational acceleration produced by a sphere. It is similar because we are dealing with a generalization of the Newtonian gravitational field. It is simpler because here we shall only calculate the rate of running of clocks (the gravitational redshift) in different places rather than the acceleration of planets that the non-uniform redshift (the curvature of time) leads to. The following steps lead to the coefficient of $(\Delta T)^2$ at any given point inside or outside the body that is the source of gravity, as long as the gravitational field is not very strong and the body is at rest or nearly so.

1. Divide the body into small pieces.
2. For each piece, multiply the density of active gravitational mass by the volume of the piece, divide by the distance to the point, and multiply by G/c^2 .
3. Add up these numbers for all the pieces of the body. Call this sum Φ .
4. The coefficient of $(\Delta T)^2$ is $-1 + 2\Phi$.

There is a name for the gravitational effect produced by this redshift factor: it is the **gravitoelectric field** of general relativity. This is the part of the gravitational field that is most like the Newtonian gravitational acceleration. The term "gravitoelectric" is used because there is a strong analogy with the electric part of the electromagnetic field, which is called the Coulomb or electrostatic field. We will encounter similar terminology below when we investigate the part of the gravitational acceleration that resembles magnetism, called the gravitomagnetic field.

The construction of the gravitoelectric field is, of course, very similar to that of Newton's field, only with a different source. It follows that, if the source of gravity is perfectly spherical, then outside the source the field is independent of the size of the region occupied by the source and is independent of whether the source is moving in and out with time, as long as it remains spherical. We saw that this was true in Newtonian gravity in Chapter 4, and it is also true in general relativity.

The gravitoelectric part of Einstein's gravity governs the motion of slowly-moving bodies, even near highly relativistic sources. The inclusion of the pressure in the source for this part of gravity leads to the most dramatic differences between the predictions of Einstein's gravity and Newton's gravity. Here is a partial list, a preview of what we will study in this and later chapters.

- *Gravitational collapse to black holes.* The large pressure of gas in relativistic objects makes the gravitational field stronger. When we study neutron stars in the next chapter, we will see that this makes it impossible for neutron stars to exist above a certain mass. Supporting the extra mass requires more pressure; that just adds to gravity and requires, in turn, more pressure. For neutron stars above a certain mass, this feedback mechanism runs away: it is never possible to add enough pressure to support the star. Instead, the star will collapse to a black hole.
- *Gravity has a magnetic-like side to its effect on matter.* We will show below that the application of the principles of special relativity to the gravitoelectric Einstein field is enough to derive the *gravitomagnetic* part of the gravitational field. The inclusion of pressure in the active gravitational mass is crucial here; it will allow us to show that the gravitomagnetic part must be present, and to derive it quantitatively, from simple arguments based on special relativity.

▷ The reason it is simpler is that the redshift is only one number at each point, while the acceleration, requiring a direction, is three numbers (a vector) at each point. If we wanted to calculate the acceleration in relativity, it would be even more complicated than the Newtonian calculation.

▷ Remember that $T = ct$ is the time measured in distance units, i.e. in light-meters.

- *Zero gravity or even anti-gravity!* As we noted right at the beginning of this book, Newton taught us that gravity is always attractive. That is because its source is mass, and mass is always positive. Well, nearly always: in quantum field theory there is the possibility of negative energy, and we will look briefly at that in Chapter 27. But in everyday situations we have come to expect gravity to be attractive. However, there is nothing unusual about **negative pressure**. In physics this is called tension. Ordinary positive pressure pushes outwards. Negative pressure simply pulls inwards. If you wrap your newspaper with a stretched rubber band, you are handling negative pressure in the rubber band. By including pressure in the source for gravity, Einstein opened the possibility that a system with very large negative pressure could have zero or even negative active gravitational mass.

The last item leads to the most dramatic differences from Newton's gravity, yet it is not always emphasized in introductions to general relativity. We will meet examples of zero and negative gravity in this book. Cosmic strings, which we first mentioned in Chapter 14, have zero active gravitational mass, their only gravitational effect being to curve space but not time. We will see what peculiar effects this can have on matter near them in Chapter 25. Even more important, Einstein himself introduced a negative-pressure field into general relativity; he called it the **cosmological constant**. His purpose was to find ways to turn the active gravitational mass negative! The cosmological constant is so important in modern physics and astronomy that we will focus on it later in this chapter. In the final chapters on cosmology we will see how scientists use negative active gravitational mass to explain the observed acceleration of the expansion of our Universe and as the basis for the theory of cosmological inflation.

Einstein's kitchen: the recipe for curving space

In this section: the active curvature mass produces the curvature when fields are weak and the gravitational field is isotropic. It works in the same way as the active gravitational mass.

We saw in the last chapter that the motion of photons past the Sun showed an extra deflection caused by the spatial curvature, and that this was a key test of general relativity. Here we learn how in general relativity this curvature is generated.

For weak gravitational fields, the coefficient in the spacetime-interval of, say, $(\Delta x)^2$, will be almost one, with a small correction. This correction is determined by the Einstein field equations. Since there are six spatial terms of the form $(\Delta x)^2$, $(\Delta x)(\Delta y)$, and so on, there are six coefficients to determine. This means that the general case will be quite complicated, possibly involving six different sources of spatial curvature.

We will simplify to just one source by assuming that the matter that produces gravity has *isotropic pressure*, which is something we defined in Chapter 7. This means that the pressure is the same in all directions, so that the averaging over directions that we used in the previous section is not necessary. This is not a strong restriction: the pressure inside a star at rest, for example, is isotropic. The pressure of the hot gases in the early Universe was similarly isotropic. So in many of the cases we will be interested in, the assumption of isotropy is fine.

Now, for weak gravitational fields in general relativity, the spatial curvature produced by matter with isotropic pressure has only *one* source, which we shall call the density of active curvature mass:

$$\text{density of active curvature mass} = \text{density of total mass} - \text{pressure}/c^2. \quad (19.2)$$

The similarity to Equation 19.1 on page 242 is striking. Both contain the mass density and the pressure, but in the case of the active gravitational mass, the pressure is multiplied by three and added, while for the active curvature mass the pressure is subtracted.

The rule for computing the coefficient of, say, $(\Delta x)^2$ is similar to the rule for computing the coefficient of $(\Delta T)^2$ from the density of active gravitational mass. The following steps will evaluate this coefficient at any given point inside or outside the body that is the source of gravity, as long as the gravitational field is not very strong:

1. Perform steps 1-3 on page 243 for the active curvature mass; call the result Ψ .
2. The coefficient of $(\Delta x)^2$ is $1 + 2\Psi$.

Because we have assumed isotropy, the coefficients of $(\Delta y)^2$ and $(\Delta z)^2$ are the same as that of $(\Delta x)^2$, and the coefficients of the mixed terms like $(\Delta x)(\Delta y)$ are all zero.

Notice that this gives the following form for the spatial distance element of a curved spacetime whose sources have isotropic pressure:

$$\Delta\ell^2 = (1 + 2\Psi) \left[(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2 \right].$$

This is exactly the form we assumed in our discussion of light deflection, Equation 18.15 on page 235, provided we identify $\Psi = GM/c^2r$. This is equal to the Newtonian field Φ , as we noted there. Now we can see that Einstein's equations do predict that $\Psi = \Phi$. They are the same because, to a good first approximation, their sources are the same. Both the active gravitational mass and the active curvature mass are dominated by the mass density ρ for systems with weak gravitational fields and small pressures. They are therefore equal, to within the accuracy with which we needed them in order to calculate the curvature for a photon's trajectory.

We have proved that Einstein's equations really do predict the observed deflection of light. We have established that general relativity passes a key observational test of its validity.

Although the active curvature mass creates the spatial curvature of Einstein's gravity, it does not lead to the kind of dramatic consequences that we saw for the active gravitational mass. Scientists generally believe that p/c^2 will not exceed in absolute value the mass-density of any relativistic field, so the active curvature mass should always be non-negative.

Einstein's kitchen: the recipe for gravitomagnetism

We are now in a position to understand one of the most remarkable features of general relativity: the existence of gravitational effects analogous to the magnetic effects of electromagnetism. When the gravitomagnetic gravitational field is present, the gravitational acceleration of a particle depends on its velocity as well as its location.

The existence of gravitomagnetism must be related to special relativity: since a particle being accelerated by gravity can be at rest with respect to one observer but moving with respect to another, the gravitomagnetic effects seen by the second observer must somehow be part of the gravitoelectric field seen by the first. In this section we will deduce the gravitomagnetic field in exactly this manner, by looking at the acceleration of a particle from the point of view of two different observers and insisting that the accelerations they predict should be the same.

It is well-known to theoretical physicists that one can deduce the existence of the magnetic field of electromagnetism from the electric field by such an argument. A particularly elegant demonstration of this was given by the brilliant American theoretical physicist Richard P Feynman (1918–1988), in which he showed how to calculate the force of magnetism by applying the rules of special relativity to the

In this section: by demanding that our theory of gravity predict the same things when used by two different observers, we show that there must be a third kind of gravitational effect, which is called gravitomagnetism. Its source is momentum and it affects only moving bodies. Our derivation follows a similar derivation of magnetism from electricity and special relativity by the physicist R P Feynman.

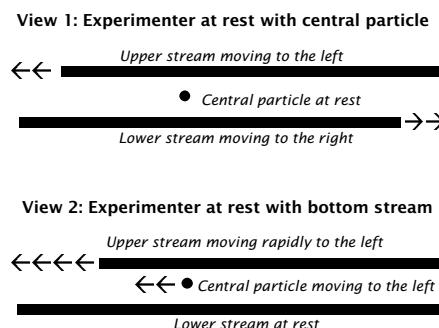
▷ Some scientists use the term *magnetogravity* instead of *gravitomagnetism*.

▷ Feynman's argument appears in his influential undergraduate physics textbook, *The Feynman Lectures on Physics*, R P Feynman, R B Leighton, & M Sands (Addison Wesley, Reading, Mass, 1964), vol 2.

Figure 19.2. Two streams of particles moving in opposite directions leave a central particle undisturbed (top panel). When viewed by an experimenter at rest with respect to the lower stream (bottom panel), the top stream has more mass and should pull on the central particle. The particle can only remain at rest if there is a velocity-dependent gravitational repulsion from a moving stream.

electric force. In this section we adapt Feynman's argument in order to derive gravitomagnetism from the gravitoelectric field. The outline and logic of the argument is presented in the main text of the section, and should be accessible to all readers. Some of the details of the calculation are reserved for Investigation 19.1 on page 250 for those who want to follow the whole calculation.

Consider the following system, illustrated in Figure 19.2. There are two streams of identical moving particles, each stream being perfectly straight and very long compared to their separation $2d$ from each other. They are also very thin (in cross-section) compared to their separation. The two streams are parallel to one another, and they have equal and opposite velocities (v and $-v$). To keep things simple, we suppose that all the particles move with the velocity of their stream: there are no random motions. The particles in the stream are so numerous that the stream is essentially a continuous string of matter.



gravitational force on the particle must be zero. Regardless of what theory of gravity one uses to calculate the force, the fact that one stream is the mirror image of the other means that the influence of one will cancel that of the other.

The particle will remain at rest in this (unstable) equilibrium position.

Let us view the same system from a rocket ship (or other experimental laboratory) that is moving at the same speed as the bottom stream. From the point of view of this experimenter, the bottom stream consists of particles that are at rest; the top stream, on the other hand, is moving to the left at speed $2v$. (We will assume that the speed v is much less than the speed of light c , so that we don't have to worry about relativistic corrections to the velocity-addition rule.) The particle in the middle is moving at speed v to the left. But the physical system is the same, just viewed by a different observer. That means that the particle in the middle remains in the middle, moving at its constant speed v , but not falling toward one or the other stream.

However, when the moving experimenter tries to calculate the gravitational forces he expects on the particle, then it is not so clear that they will balance. Let us consider what this experimenter expects to happen if he believes Newtonian gravity is all he needs to know, so he does not use Einstein's active gravitational mass. He assumes that the total mass-energy of the stream creates gravity, and he (like the first experimenter) has a machine that can measure accurately the mass per unit length along each stream. We will compare the mass per unit length that the two experimenters measure for each stream.

In Newtonian gravity, if the stream has a very small cross-section, then the gravitational acceleration produced by the string will depend only on the amount of (rest) mass in the string per unit length. We call this number μ , and it has units of kilograms per meter.

Now we do the following simple idealized experiment.

We place a particle exactly in the middle between the two streams, at rest. Because of the symmetry of the situation, the net

The top stream is moving *faster* relative to the second (rocket) experimenter than to the first. This means that each particle will have a larger mass-energy as measured by the second experimenter. What is more, the Lorentz–Fitzgerald contraction of lengths pushes the density of mass-energy higher still. So the second experimenter measures a much *higher mass-energy* per unit length for the top stream than the first experimenter does.

The bottom stream, by contrast, is at rest with respect to this second experimenter, while it is moving relative to the first experimenter. The situation is the reverse of that for the top stream, so the same reasoning leads to the conclusion that the second experimenter measures a *smaller mass-energy* per unit length for the bottom stream than the first experimenter does.

The rocket experimenter, assuming as he does that Newtonian gravity is the whole story, expects that the gravitational force exerted by the top stream will be larger than that exerted by the bottom one, and that the particle in the middle will begin to fall towards the top stream.

But the particle does not fall toward the top stream. So the experimenter in the rocket must conclude that there is more to gravity than the simple Newtonian force.

We knew already on general grounds that Newtonian gravity is not compatible with special relativity. Here we have an explicit demonstration of it.

Now, we already know about one correction to Newtonian gravity: the active gravitational mass in relativity includes the pressure term that is shown in Equation 19.1 on page 242. So maybe all we have to do is calculate what we have called the gravitoelectric gravitational acceleration in general relativity; maybe then the central particle will experience exactly balancing accelerations.

What, however, is the pressure term in our situation? Our streams have no conventional fluid pressure, since the particles have no random motions. However, the overall bulk motion of the particles creates what physicists call a “ram pressure”, which is basically the pressure that the stream would exert if it were running up against (ramming into) a wall. This pressure does contribute to the gravitational field. We calculate the ram pressure in Investigation 19.1 on page 250, and we find it is proportional to μv^2 . When we put that into the expression for the active gravitational mass, however, it does *not* correct the imbalance of the accelerations as computed by the rocket experimenter. In fact, it is easy to see that it makes it worse. The pressure is positive, and it is non-zero only in the top stream, where the gravitational attraction was already too high.

Adding in pressure only makes the imbalance worse! We are driven to the conclusion that there has to be more to gravity in general relativity than just the gravitoelectric part of the field.

We clearly need a further acceleration, produced by the top stream, that *repels* the central particle, keeping it in equilibrium with the weaker gravitoelectric acceleration produced by the bottom stream. Since we did not need to invoke such an acceleration when we looked at these streams from the point of view of an experimenter at rest with respect to the central particle, it is natural to expect that this acceleration will turn out to be associated with the *motion* of the particle with respect to the second experimenter. In order to cancel out the excess active gravitational mass density of the upper stream, this velocity-dependent force must work in such a way that a moving stream *repels* a particle that is moving in the same direction.

► You might wonder what would happen if, despite our earlier objections, the moving experimenter assumed that *rest-mass* created Newtonian gravity, instead of total mass-energy. It should be easy for you to see that the Lorentz–Fitzgerald contraction still makes the rest-mass density higher in the top stream than the bottom, so this variant of the assumed Newtonian force would also draw the particle towards the top.

► In fact, there is an alternative relativistic theory of gravity in which the pressure enters the active gravitational mass with the opposite sign, and with the right coefficient so that the ram pressure cancels the higher density of the upper stream and leaves the particle in balance, with no further corrections. This theory has no gravitomagnetism. However, it predicts the wrong result for the anomalous perihelion shift of Mercury, even giving it the wrong sign, so it is not a viable theory. This illustrates the important point that the ultimate test of a theory is its agreement with experiment.

The situation is very close to that in magnetism: an electric current (stream of moving positive charges) will create a magnetic field that actually *attracts* a positively charged particle that is moving in the same direction as the current. Here we see that a moving stream of particles will repel a particle moving in the same direction. The sign of the effect is different (attraction in one case, repulsion in another), but this is just because in electromagnetism the sign of the electric part of the acceleration is also different from that in gravity: electric charges of the same type repel each other, while in gravity two masses attract. Apart from this sign, there is such a close analogy to magnetism that we call this velocity-dependent gravitational effect *gravitomagnetism*.

►It should be clear from our derivation that the words "gravitoelectric" and "gravitomagnetic" are used only to draw an analogy with electromagnetism. They are purely gravitational effects; they have their source in the mass and momentum of particles, not in electric charge or electric current.

The gravitomagnetic effect is created by the moving stream, so it is a gravitational effect whose source is the *momentum* of the particles. We have therefore found the field created by the third source of gravity in the list on page 240. In the first experimenter's view, both streams create gravitomagnetism, but the central particle is at rest, so it does not feel the effect. In the second experimenter's view, the bottom stream is at rest and therefore does not create this effect, but the top stream does, and it just compensates the extra gravitational attractiveness of the top stream to produce the same net gravitational attraction as the bottom stream exerts. We calculate the size of this effect in Investigation 19.1 on page 250. The argument gives exactly the gravitomagnetic effect that one could calculate from Einstein's field equations directly, with the mathematics of differential geometry! Our derivation is just as good, and uses only elementary algebra.

By analogy with the magnetic field, the direction of the gravitomagnetic effect can be determined by something we might call the two-hand rule, as follows. Let the thumb of your right hand point in the direction of the momentum of the top stream. Then let your fingers curl up around this direction. The fingers follow lines of the gravitomagnetic part of the gravitational field, which are circles around the stream. Now, to calculate the effect on a passing particle, take your *left* hand and let the fingers curl in the following way. First point the fingers in the direction of the motion of the passing particle. Then curl them so that their tips point along the direction of the gravitomagnetic effect, which you just determined using your right hand. When your left hand is oriented so that the fingers can curl from the one direction to the other as described, then your left thumb points in the direction of the gravitomagnetic acceleration of the particle.

The size of the gravitomagnetic effect, as calculated in Investigation 19.1 on page 250, has a simple formula, expressed as a correction to the gravitoelectric acceleration produced by any source. If the Newtonian gravitational field of a system would produce an acceleration that has magnitude a_N , and if the source moves with speed v_s and the particle with speed v_p , then the magnetic-type gravitational acceleration of the particle will have magnitude

$$a_M = a_N \frac{4v_s v_p}{c^2}. \quad (19.3)$$

The direction of the magnetic-type acceleration is given by the two-hand rule.

This equation allows us to compute the gravitomagnetic acceleration produced by *any* moving system on a moving particle, if we know the two velocities and the Newtonian acceleration the system produces.

The idealized, infinitely long streams of particles have served their purpose, so we can forget them now and focus on more realistic systems.

For example, let us write down the magnitude of the gravitomagnetic acceleration due to a single particle source of mass M moving with speed v_s . If we look at

►We can use the Newtonian acceleration a_N here rather than the full gravitoelectric acceleration, because the extra pressure terms are already corrections to a_N of order v^2/c^2 , so they become terms of order v^4/c^4 in a_M , and we have neglected such corrections in this argument.

the acceleration at a distance r from it, then the Newtonian acceleration is, in magnitude, $a_N = GM/r^2$, so the gravitomagnetic acceleration on a particle with speed v_p has magnitude

$$a_M = \frac{4GMv_s v_p}{c^2 r^2}. \quad (19.4)$$

Like the Newtonian acceleration, it falls off as $1/r^2$.

Notice that we can write the gravitomagnetic force on a particle of mass m , which is just ma_M , in the form

$$F_M = \frac{4G}{c^2 r^2} (Mv_s)(mv_p).$$

It is therefore possible to regard gravitomagnetism as a coupling between the *momentum* of the source and that of the particle.

In this way we see that momentum creates its own kind of gravity.

The geometry of gravitomagnetism

So far, we have talked in Newtonian language about gravitomagnetism, describing the way it acts on particles. It is natural to ask how it fits into the geometrical picture: where, in the calculation of the spacetime-interval, does gravitomagnetism come in?

First we have to decide what we expect to find in the spacetime-interval. The spacetime-interval represents the gravitational field created by the source of gravity. It does not contain any properties of the particles that are affected by gravity: they move on geodesics of this spacetime geometry. So when we look for the gravitomagnetic acceleration terms in the spacetime-interval, we are looking only for the first factor in the following equation, which is just Equation 19.3 re-written in a convenient way:

$$a_M = \left(a_N \frac{4v_s}{c} \right) \frac{v_p}{c}.$$

We have factored out the part of the acceleration that depends on the particle, its (dimensionless) speed v_p/c , so that what is inside the large parentheses is the part due just to the source of the field. This is the *gravitomagnetic field*:

$$\text{gravitomagnetic field} = \frac{4v_s}{c} \times \text{Newtonian gravitational field.} \quad (19.5)$$

The rule is that the magnitude of the gravitomagnetic effect on a particle is just the gravitomagnetic field times the dimensionless velocity v_p/c of the particle. The direction of the effect is given by the two-hand rule.

We expect to find the gravitomagnetic field of Equation 19.5 encoded somewhere in the spacetime-interval. We can discover where it is by the application of a symmetry argument. Consider what happens to our example when we reverse the sense of time, as if we took a video of the experiment and played it backwards. Then all the velocities would go the other way, and by the two-hand rule the sense of the gravitomagnetic field would reverse. The end effect, the acceleration of the particle, would not change: it would still be repelled from the top stream. But this would come about because of two compensating changes of sign. We would be multiplying the particle's own velocity, which has changed sign, by the gravitomagnetic field, which has also changed sign.

So the gravitomagnetic field itself must be contained in the spacetime-interval in a term that changes sign when we change the sign of T . Moreover, we get a similar change of sign if we reflect the experiment in a mirror perpendicular to the

In this section: gravitomagnetism comes from the mixed time-space coefficients in the general interval.

▷ Just where we place the factors of c in defining these parts of Equation 19.3 is, of course, arbitrary, but it seems simplest to keep things dimensionless where possible, so that the gravitomagnetic field has the same dimensions as the Newtonian field.

Investigation 19.1. How big is gravitomagnetism?

Here we shall see how to calculate the gravitoelectric part of the gravitational attraction, which comes from the active gravitational mass in general relativity, and we will find the shortfall that must be made up by gravitomagnetism. The first experimenter sees a symmetrical situation, so it is not of interest to us to calculate the forces from his point of view: they will only cancel out completely and leave the particle at rest. So we will focus on the second experimenter, flying in a rocket that is moving at the same speed as the bottom stream.

The bottom stream produces the same gravitoelectric acceleration in general relativity as it would in Newtonian theory, because it has no pressure and no velocity (as measured by the second experimenter). Let us call the mass per unit length of this stream as measured by the second experimenter μ' . This differs from μ , which is measured by the first experimenter, but we won't need to find the relation between the two.

In Newtonian theory, an infinitely long line with a mass-per-unit-length of μ' will create a certain gravitational acceleration in a particle a distance d away. This acceleration is proportional to μ' . It also depends on d , but since in our situation d will not change, and it is the same for both streams and (importantly) for both experimenters, we won't need to know the dependence on d . We will just write the acceleration produced by the bottom stream as measured by the second experimenter as

$$a_N = \alpha\mu',$$

where the constant α contains all the things we don't want to bother with.

In general relativity, the gravitoelectric acceleration produced by the *upper* stream, as calculated by the second experimenter, will be different from the Newtonian acceleration we have just written down for three reasons. First, the second experimenter will measure the mass of each particle to be larger than its rest-mass, because of the extra kinetic energy. Second, the experimenter will measure a smaller separation between the particles, because of the Lorentz-Fitzgerald contraction. And third, general relativity tells us that the pressure has to be added into the expression for the active gravitational mass, as in Equation 19.1 on page 242. We need to work out all three corrections.

- The transformation of mass.* We learned from Equation 15.6 on page 190 how mass depends on speed. Given that the speed of the top stream is $2v$, the mass of each particle as measured by the second experimenter is a factor $[1 - (2v)^2/c^2]^{-1/2}$ larger than the rest-mass.
- The Lorentz-Fitzgerald contraction.* By Equation 15.5 on page 188 for the Lorentz-Fitzgerald contraction, the particles in the top stream are closer to one another by the factor $[1 - (2v)^2/c^2]^{-1/2}$, which further raises the density of mass along the stream. Thus, to the first experimenter, the mass density per unit length of each stream is

$$\text{mass density of moving stream} = \mu / [1 - (2v)^2/c^2].$$

We will work only with first corrections to the Newtonian formulas, so we can write (recall Equation 5.2 on page 43)

$$[1 - (2v)^2/c^2]^{-1} \approx 1 + (2v)^2/c^2 = 1 + 4v^2/c^2.$$

This implies

$$\text{density of top stream} \approx \mu' + 4\mu' v^2/c^2.$$

- Pressure contribution to the active gravitational mass.* In our example there is no ordinary pressure inside the streams, but there is still an effect of the same type. If the stream were

to run into a solid wall, like spraying a jet of water from a hose against a wall, then there would be a large pressure on the wall, even if there were no internal pressure at all in the stream. This is called the "ram pressure" of the stream, and in general relativity, this kind of ram pressure will create gravity the way ordinary pressure does.

The ram pressure can be calculated by asking what sort of pressure the wall has to exert on the water stream in order to avoid being knocked over by the spray. If the density of the stream (this time we mean mass per unit volume) is ρ and its speed is u (which we will set to $2v$ later), and if the cross-sectional area of the hose is A , then in a small time t the mass of water that hits the wall will be $m = \rho \times A \times ut$. This water has momentum mu . To stop the water, the wall has to exert a force equal to the change it makes in the momentum of the water divided by the time, or $F = \rho Au^2$. The force per unit area, or in other words the pressure exerted by the wall as it continuously resists the water stream, is $p_{\text{wall}} = \rho u^2$. This is, by Newton's law of action and reaction, the same as the ram pressure of the water stream itself, or of any other stream of uniformly moving particles.

Now, in our example we are interested in the active gravitational mass per unit length of our streams, not their mass per unit volume. This is because we assume the streams are so thin that all the particles in a given cross-section of the stream are effectively the same distance from the central particle. Since the ram pressure is just u^2 times the mass density ρ , then by analogy the ram pressure contribution to the active gravitational mass per unit length will be $\mu' u^2/c^2 = \mu' (2v)^2/c^2 = 4\mu' v^2/c^2$.

We have only to worry about the requirement in Equation 19.1 on page 242 that we need the *average* pressure. The pressure in the stream in directions perpendicular to the stream is zero: there is no ordinary pressure and there is no velocity to make a ram pressure. So the average over the three directions of the pressure is $(4\mu' v^2 + 0 + 0)/3 = 4\mu' v^2/3$. Then the final correction is three times this divided by c^2 :

$$\begin{aligned} & \text{pressure part of the active gravitational mass} \\ & = 4\mu' v^2/c^2. \end{aligned}$$

This adds to the density of the top stream to give

$$\begin{aligned} & \text{total active gravitational mass} \\ & \approx \mu' + 4\mu' v^2/c^2 + 4\mu' v^2/c^2 = \mu' + 8\mu' v^2/c^2. \end{aligned}$$

When all three corrections are added into the active gravitational mass, we find that the gravitoelectric Einstein gravitational acceleration of the particle due to the top stream, as measured by the experimenter at rest with respect to the particle, is

$$a_E = \alpha\mu' (1 + 8v^2/c^2).$$

We find, therefore, that the rocket experimenter calculates that the gravitoelectric gravitational attraction of the top stream exceeds that of the bottom by $8(v^2/c^2)\alpha\mu$. If there were no other gravitational effects, the central particle would move upwards with this acceleration.

Since the central particle does not move, there must be a magnetic-type acceleration, depending on velocities, that exactly compensates this. Since it should depend on both the speed of the particle, $v_p = v$, and the speed of the source, $v_s = 2v$, we can write this acceleration as a repulsion from the source of magnitude

$$a_M = 4(v_p/c)(v_s/c)\alpha\mu.$$

x -direction. This changes the direction of all the velocities in the same way that changing the sense of time did. So again, the gravitomagnetic effect must be in a term that changes sign when we replace x by $-x$. There is only one term that changes sign when we do either operation, and that is the mixed term containing the product $(\Delta T)(\Delta x)$.

The coefficients of the mixed terms between time and space coordinates in the spacetime-interval, like $(\Delta T)(\Delta x)$, create the gravitomagnetic effects in the motion of particles following geodesics of a geometric gravitational field.

Gyroscopes, Lense, Thirring, and Mach

Once we realize that gravitomagnetism exists, we can find many situations where it can be seen. The most important are the effects caused by rotating masses. Long streams of particles, such as we treated in the last section, are rare in the Universe, but rotating stars and black holes are common. In this section we will see how to estimate the gravitomagnetic effect of a rotating star, how this is being measured today for the Earth, and how it is related to an old philosophical idea called Mach's principle.

From the two-hand rule, we can determine the effect of gravitomagnetism on bodies moving in other ways than the central particle of the example we studied first. In particular, suppose in Figure 19.2 on page 246 that a particle is moving directly *towards* the top stream, from above it. Then the gravitomagnetic effect will bend its path *in the direction of motion of the stream*. By symmetry, this will happen to any particle approaching the stream. This effect has acquired a rather dramatic name in general relativity. It is called the **dragging of inertial frames**. What this means is that the stream seems to change the local standard of rest. Particles that are at rest far away and then fall toward the stream are dragged along it a little, as if the stream were a jet of water pushing though still water, entraining some of its surroundings along it.

Let us see how this dragging can be important in realistic situations. Consider the gravitational field near the rotating Earth. The rotation of the Earth can be approximated, for the purposes of our little discussion, as a stream of matter moving around a loop. Then if we are near one side of the loop, we see gravitomagnetic forces from the near side, pushing us one way, and from the far side, pushing us the other way. These effects tend to cancel each other, but since the force depends on distance, the cancellation is not perfect, and the near side wins. The further we are from the Earth, however, the less significant is the difference between the distances to the two sides of the loop, so the better is their cancellation.

The net result is that the gravitomagnetic force caused by the rotation of the Earth falls off with distance from the Earth faster than it does from a single moving particle, proportional to $1/r^3$ rather than the $1/r^2$ of the basic Newtonian force.

The rotation of the Earth also gives this dragging force a twisting character. This is most easily seen if we imagine placing a spinning gyroscope exactly in the *center* of the rotating loop that we take as a model for Earth (Figure 19.3 on the next page). This idealized experiment will help us understand what will happen in more realistic situations.

Suppose the gyroscope is oriented horizontally, with its axis parallel to the equatorial plane, pointing momentarily at longitude 0° . Suppose also that the gyro is spinning in a positive sense, which means it is spinning counterclockwise when looking down the axis from the longitude 0° point on the loop.

In this section: through gravitomagnetism, spinning bodies can cause particles near them to rotate in the same sense. Two experiments are trying to measure the effect using gyroscopes and satellites. The effect is as close as general relativity comes to supporting the ideas of Mach on where inertia comes from.

►In relativity, a *frame* is the coordinate system of an observer, so the term "dragging of inertial frames" describes the way a freely-falling observer is swept along in the direction of rotation by gravitomagnetism.

Figure 19.3. Idealization of the geometry of a spinning Earth dragging a gyroscope situated at its center. We represent the Earth as a loop of mass concentrated at the equator, and the gyroscope as a disk spinning about a horizontal axis pointing toward longitude 0° .

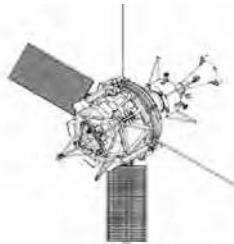
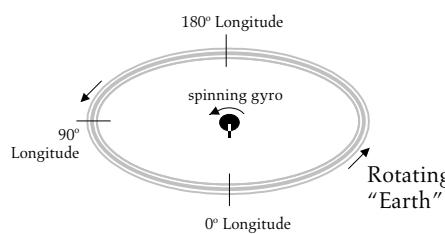


Figure 19.4. Drawing of the Gravity Probe B (GP-B) satellite, which will carry the most sensitive gyroscope ever constructed into orbit to measure the Lense–Thirring effect. The gyroscope will change its direction by only 42 milli-arcseconds in one year, and GP-B is designed to measure that to an accuracy of 1%. This angular precision, half a milli-arcsecond, is about the angular size of a medium-sized dog at the distance of the Moon!

Drawing courtesy of Gravity Probe B.



Figure 19.5. This fused-quartz sphere coated with superconducting niobium is one of the gyroscopes carried on GP-B. The size of a table-tennis ball, it is so spherical that its irregularities are nowhere more than 40 atoms high. This is just one of many challenges that have been met in designing this extraordinarily sensitive satellite.

Image courtesy of Gravity Probe B.

Now, the gyro is just a loop of mass moving in a circle about its own axis. On the top of the gyro the mass of this loop is moving towards longitude 90° W (just off the coast of Ecuador). The part of the Earth in western longitudes is moving eastward, and exerts a dragging force on this part of the gyro to pull it towards the direction of 0° longitude.

longitude. The part of the Earth on the other side of the gyro, in eastern longitudes, is moving in the opposite direction, but the gyro is moving away from it, so it "anti-drags" the gyro, again pulling the top part of it towards 0° longitude. The bottom part of the gyro is moving in the opposite sense, so it must feel a force towards 180° longitude, in the South Pacific. The net result of these two forces is a torque (twisting force) trying to pull down on the part of the gyro's axis that points toward 0° and up on the opposite side.

Now, we know what happens when we try to do this to a gyro: it rebels. It simply turns to the side.

The effect of the gravitomagnetic forces, therefore, will be to change the direction of the gyro's axis, causing it to rotate slowly in the same direction as the Earth spins. It is not hard to convince oneself that this will happen to any gyro oriented horizontally, even if it is not at the center of the Earth. This is called the Lense–Thirring effect, after the two scientists – Josef Lense (1890–1985) and Hans Thirring (1888–1976) – who discovered it only two years after Einstein published his general theory.

A gyro can be used to measure the Lense–Thirring effect. Two experiments are presently underway using very different kinds of gyros. One of them is called Gravity Probe B, illustrated in Figure 19.4. This is planned for launch by NASA in 2003; it will carry a very sensitive gyroscope (Figure 19.5) into orbit around the Earth to try to measure the predicted effect over a period of a year or more. The other experiment uses existing satellites, called LAGEOS and LAGEOS2, shown in Figure 19.6. If a satellite has an orbit that goes over the Poles, it is moving in the same way that the mass of the gyroscope was moving in our example above, so the orbit will get twisted by the Lense–Thirring effect in the same way: the orbit will gradually precess eastwards. By precise tracking of the orbits of these satellites, which have been specially designed to minimize the effects of atmospheric drag, and which have nearly polar orbits, the second group of scientists is presently measuring the dragging. Their initial results have confirmed the predictions of Einstein's theory, and higher accuracy is expected in the near future.

Astronomers are beginning to see the effects of frame-dragging near neutron stars and black holes. We will return to the evidence for this in Chapter 21, but it seems that the Lense–Thirring effect may soon be used to measure the spin of black holes in astronomical systems. Other astronomers are proposing a very high-accuracy astrometry satellite called GAIA, a successor to the Hipparcos mission described in Chapter 9. This would be able to see the extra deflection effects produced by dragging as light passes near the Sun. The Sun is spinning, so light that passes on one side of it will be affected by dragging differently than light on the other side, and it will be possible to measure the interior spin of the Sun for the first time in this way.

Suppose the experiments near the Earth are not able to confirm the details of frame-dragging: suppose gravitomagnetism is not as predicted by general relativity. What then? Our derivation makes it clear that only two assumptions are needed to derive the standard formulas: special relativity and the Einstein expression for the active gravitational mass. We would not like to give up special relativity, since it is tested in many other places. We would therefore have to look for a different expression for the active gravitational mass, and that would have all kinds of implications for gravity.

Since Einstein's equations work so well in other situations, it seems very likely that the Gravity Probe B and LAGEOS experiments will verify the Lense-Thirring effect. But surprises are always possible!

The dragging effects of gravity are reminiscent of philosophical ideas that go back to the Moravian physicist Ernst Mach (1838–1916). Mach was intrigued by the question of why bodies have inertia. Why does it require a force to accelerate a mass? What is so special about the state of uniform motion that it requires no acceleration? Put simply, uniform with respect to what? Mach suggested that the Universe itself establishes what is meant by uniform velocity, that a velocity can be maintained without an external force if it is uniform with respect the Universe. He speculated that this condition had a real cause, that bodies exerted an influence on one another that resisted their relative acceleration. Although Mach did not turn these ideas into a successful theory, they appealed to Einstein and he gave them as one of the influences that shaped the way he searched for a relativistic theory of gravity.

The dragging of inertial frames seems Machian in spirit. It is a real influence that seems to try to bring one thing closer to the state of motion of another. But the fit to Mach's ideas is very imperfect. For one thing, gravitomagnetism depends on the direction of motion. A body falling towards a stream of matter is indeed pulled in the direction of its motion, but a body moving away from the stream is accelerated in a direction *opposite* to the motion of the stream! And a body at rest feels no influence from the motion of the stream at all.

In fact, despite Einstein's interest in Mach's ideas, Einstein's own theory sheds no light on what creates inertia.

The cosmological constant: making use of negative pressure

The history of the cosmological constant is one of the oddest chapters in the development of general relativity. Einstein reluctantly and belatedly introduced this new term into his theory, and then he later withdrew it. But now astronomers think they have measured it, cosmologists imitate it in their theory of cosmological inflation, and physicists find that it comes naturally out of their theories of high-energy physics. In this section we will simply describe the way the cosmological constant works, and what is special about it. We are thereby preparing ourselves for our discussion of the physics and astronomy of the cosmological constant in the last chapters.

When Einstein invented general relativity, astronomers did not yet know that the Universe was expanding. Einstein wanted to be able to make a mathematical model of the whole Universe that was static, neither expanding nor contracting. To do this he needed something that would counteract the attractive force of the matter in the Universe. Fortunately for him, his equations gave him the right loophole: negative pressure. If he could introduce enough negative pressure, then he could arrange for the total density of active gravitational mass to be zero.



Figure 19.6. The LAGEOS satellite is the complete opposite of GP-B. It is a passive satellite, with no working components. Covered in mirrors, its only job is to reflect laser beams back to Earth, which are used in range-finding to find its exact position. The relativity experiment is a spin-off from the satellite's main mission, which is to track continental drift by measuring the motion of the ground stations that track the satellites. From the tracking data, however, scientists are beginning to discern the gradual precession of the orbit of one of them induced by the Lense-Thirring effect. Image courtesy GSFC.

In this section: the cosmological constant can be viewed as a physical fluid with a positive density and a negative pressure. We derive the remarkable and unique properties of this special fluid: it has no inertia, exerts no pressure forces, stays the same density when it expands or contracts, and creates a repulsive gravitational field: anti-gravity. These properties allowed Einstein to introduce it safely into his equations in order to stop the Universe from collapsing.

As we remarked above, negative pressure is called tension. But in ordinary materials, it normally is present only if the material is acted upon by outside forces, such as being stretched in one direction like a rubber band. Normal materials do not have tension in their resting state.



Figure 19.7. Ernst Mach is mainly remembered today for his work on supersonic motion: the Mach number of a projectile or aircraft is the ratio of its speed to the speed of sound. But he also speculated about profound questions in physics, psychology and philosophy, often advocating his positions stubbornly. He was one of the most vocal (and last!) opponents of the atomic theory of matter, and strongly attacked Boltzmann's theories, despite the fact that he and Boltzmann were colleagues at the University of Vienna. Regarding inertia, Mach was dissatisfied that physicists since Newton had studied only the forces that were required to accelerate masses, but not why the masses had inertial mass in the first place. Mach hoped to find a deeper physical principle underlying Newton's laws. Image courtesy Charles University, Prague.

No ordinary matter displays isotropic negative pressure in its normal state. Einstein's suggestion was something entirely new. He had no physical model for his cosmological negative pressure. It was a mathematical device to produce a Universe with zero gravitoelectric force on the large scale.

It is not hard to imagine why Einstein was never happy with this idea, despite the fact that it did what he wanted. For one thing, it was what scientists describe as *ad hoc*, something that has no other justification than to patch things up. The negative pressure had no foundation in observation, and was introduced simply to rescue the theory from the difficulty of the expanding or contracting Universe that it predicted. Einstein had no physical mechanism for producing the pressure.

Even worse, in order to make the Universe static, the amount of pressure had to be exactly right, just enough to cancel out the energy density of the universe in the active gravitational mass. If the pressure were not large enough to cancel the attraction of the energy density, then the Universe would slow down and perhaps re-collapse; if the pressure over-compensated for the energy of the Universe, then the Universe would expand in an accelerated way. Einstein's static universe model was *unstable* to small changes in its density.

But Einstein recognised that negative pressure was the only way he could get general relativity to give a static Universe, so he pursued the idea.

Einstein found that there was one and only one way to introduce this negative pressure and still preserve his principle of relativity. His cosmological constant introduces an energy density and pressure into the Universe that are constant in time and in space, and that moreover are the same no matter which observer measures them. The cosmological "fluid" is completely invariant. This brilliant mathematical insight has consequences in modern cosmology well beyond anything Einstein could have imagined.

Let us see what we can make of this idea.

Einstein wanted to introduce negative pressure without giving up the most fundamental feature of general relativity, that it does not pick out any special observer, or place, or time. Since his cosmological negative pressure was to be fundamental, not tied to any accidental matter field or configuration, he needed the pressure to be constant in space and in time, so that an observer could not pick out any special place or time by measuring the pressure. It had to be a fundamental *constant* of nature. Einstein actually introduced, instead of a fundamental pressure, a fundamental constant Λ , which he called the *cosmological constant*. The uniform cosmological pressure he needed, p_Λ , is defined in terms of Λ by

$$p_\Lambda = -\frac{c^2 \Lambda}{8\pi G}. \quad (19.6)$$

The sign allows Λ to be positive to give the negative pressure required for a static universe. The other constants in the definition show that Λ itself has the dimensions of a frequency squared, or $1/(time)^2$.

Einstein found that he could make the pressure invariant if the cosmological constant also generated a mass density

$$\rho_\Lambda = -\frac{p_\Lambda}{c^2} = \frac{\Lambda}{8\pi G}. \quad (19.7)$$

The cosmological fluid that has this pressure and density has remarkable properties. First, let us ask if there is any way we can detect this fluid, other than by observing its gravitational effects. What are its local properties?

- *Zero inertial mass density.* One might ask if this cosmological fluid could be felt in non-gravitational ways. For example, does it have inertia, so that it would make things harder to move? We calculated the inertial mass density in Chapter 15, and we saw in Equation 15.10 on page 193 that it is $\rho + p/c^2$. With $p_\Lambda = -\rho_\Lambda c^2$, this fluid has *zero* inertial mass! It can be accelerated with no cost, no effort. This property is the key, as we will see momentarily, to the invariance of the pressure and density against a change of observer.
- *Zero pressure force.* With a large negative pressure, surely this fluid would exert observable pressure forces on things in the Universe. But no, pressure forces act only through pressure *differences*, as we saw in Chapter 7. A uniform pressure, even a negative one, exerts no force.

The cosmological fluid is remarkable indeed:

Einstein's cosmological constant is *undetectable* in non-gravitational experiments. It contributes nothing to local dynamics. It offers no resistance to objects moving through the vacuum. Its pressure is uniform, so it exerts no direct forces on objects. You can't *feel* the cosmological energy density or pressure. The vacuum is just as empty with it as without it, except for its gravitational effects.

This aspect of the cosmological constant was particularly repugnant to Einstein, who had only recently succeeded in getting rid of the nineteenth-century notion of the ether, as we saw in Chapter 15. Now apparently he was forced to introduce something just as strange.

Now let us see how this fluid could have these invariant properties. Normally, the density and pressure of a fluid depend on the observer, on the speed of the fluid relative to the observer. Consider, first, how the pressure of the cosmological fluid might depend on its motion. We saw in our earlier derivation of gravitomagnetism, in our discussion of the active gravitational mass of a stream of particles, that when a fluid moves, then in this direction the density contributes something to the pressure. We called this the ram pressure, and saw that it equals ρv^2 . But that discussion assumed that the fluid was non-relativistic, in the sense that not only was the speed v small but also the pressure was small. When the pressure is large, so that p/c^2 is similar in size to ρ , then the ram pressure must be modified. We must use the inertial mass per unit volume, $\rho + p/c^2$, instead of just ρ . The inertial mass per unit volume is the quantity that measures the inertia of the fluid, which determines the pressure it would exert if it ran into a wall. So the ram pressure of a relativistic fluid moving at a small speed v is $(\rho + p/c^2)v^2$. Now, we have already seen that the inertial mass density of the cosmological-constant fluid is zero. Therefore, its pressure as measured by an observer who is moving with respect to the fluid is exactly the same as for an observer at rest. The pressure p_Λ of this fluid is an invariant.

In the same way, the density ρ_Λ is also an invariant. We leave the proof of this to Investigation 19.2 on page 257. And there is one other property that Einstein

▷ As if to rub salt into Einstein's wounds, modern physicists use the term **quintessence** for some new theories of physics that introduce fields with negative pressure.

Quintessence was the name Aristotle used for the ether.

►It is worth asking how Einstein arrived at this remarkable prescription for a cosmological fluid. He certainly did not follow the route I have taken in presenting it here; this method fits well with physicists' perspective on the constant today, but it was not Einstein's perspective. In fact, he was led to Λ by the principle of general covariance. Once one has studied the full mathematics of Einstein's equations, the cosmological constant actually seems like a rather natural modification of the theory. Einstein regarded Λ as a fundamental constant of Nature, which he introduced as a modification of the field equations (i.e. of the *recipe*). He did not think of the cosmological constant itself as a fluid, as a source of gravity, as a new *ingredient* for the old recipe. We shall see this in Equation 19.9 on page 258.

In this section: we meet the full field equations of general relativity and learn why Einstein was led to postulate them, and in what sense they are simple and elegant.

►This is exactly the kind of compensation that we saw above when we used the gravitomagnetic force to balance the excess gravitoelectric force seen by a moving observer. We see from that example how well Einstein succeeded.

required. His equations of gravitation require that any matter of fluid in spacetime should obey the laws of conservation of energy and momentum in any small volume. His new cosmological fluid needs to pass this test too. We shall see, again in Investigation 19.2, that the law of conservation of energy insures that, as the Universe expands, the density and pressure of this fluid remain *constant*, just as Einstein required. With these properties, the cosmological constant provided Einstein with just what he needed: a force that could keep the Universe static and at the same time did not single out a preferred observer, place, or time.

Then came Edwin Hubble. When Einstein learned of Hubble's discovery that the Universe was expanding, he bitterly regretted having invented the cosmological constant. He reasoned that, if he had had the courage to stay with his original theory, he would then have predicted the expansion before it was discovered, and his prediction might well have led astronomers to look for the expansion earlier than they did. The expanding universe would have been seen as a further experimental test of and triumph for general relativity. To Einstein, one of his greatest blunders was not having had confidence in his original equations in the cosmological arena.

Physicists today take a more generous view of Einstein's "blunder". Spurred on by theoretical considerations in fundamental physics, which suggest that this kind of cosmological fluid could be a natural consequence of theories of high-energy physics, physicists are looking for ways to predict a cosmological constant with a value that would account for recent astronomical observations that the expansion of the Universe appears to be accelerating. They actually use the word "field" instead of "fluid" to describe the cosmological constant, but that is nomenclature. We will return to a discussion of this in our final chapter, Chapter 27.

The big picture: all the field equations

We have seen the detail of Einstein's theory, we have used it to calculate the deflection of light by the Sun, and we have shown that the gravitational field includes interactions between momenta and between spins. We could go on to study the phenomena of relativistic gravity. But we have not yet asked where Einstein's theory came from, what led Einstein to his creation. We will spend the rest of this chapter trying to look at general relativity from Einstein's own perspective.

Einstein was looking for equations that would generate the curvature that represents gravity and at the same time obey the principle of general covariance. We have seen that the way we describe geometry, for example the coefficients in the spacetime interval, depends on the coordinate system. Einstein had to find a way to allow the geometrical description and the sources to change when the observer changed, but to get the same geometry in the end. If things worked out correctly, the different parts of the gravitational field would fit together in just the right way to compensate for the different values measured by different observers for the individual sources.

This is easy to say, but hard to do. Einstein could make no assumptions to simplify the form of the spacetime interval; he had to work with the most general form. We have written such a spacetime interval for two dimensions in Equation 18.5 on page 228, and we saw it had three coefficients, which depended on three functions A , B , and C . In four dimensions, there are ten coefficients: four for the terms like $(\Delta T)^2$ and $(\Delta x)^2$, and another six for mixed terms like $(\Delta x)(\Delta y)$ and $(\Delta T)(\Delta z)$. That means

Investigation 19.2. The remarkable physical properties of the cosmological fluid

Einstein defined his cosmological constant in a very special way, ensuring that there was a strict relationship $p_\Lambda = -\rho_\Lambda c^2$. This can be thought of as the equation of state of the cosmological fluid. (We introduced the notion of an equation of state in Chapter 7.) This equation of state has a remarkable property. It guarantees that, as the Universe expands, the mass density of this cosmological “fluid” remains *constant* in time. Unlike all normal gases, a fluid with $p = -\rho$ does not get diluted by expansion. Nor does its density depend on its speed, so it does not pick out any preferred observer.

To see how the density remains constant, consider an isolated box filled with such a fluid. Suppose the box has initial volume V and then expands slowly to $2V$. Since the fluid in the box has tension, a force is required to expand the box. The force does work, as defined in Equation 6.20 on page 62, and this adds energy to the fluid. The mass equivalent to this energy adds to the mass already in the box. We shall show that when the pressure is that of this peculiar cosmological fluid, the added energy is just enough to insure that the density of the fluid in the larger volume is the same as in the smaller.

We can make this verbal explanation quantitative with a simple set of calculations. As our first step we will find out how the energy in any fluid changes when it expands or contracts. Consider a rectangular box of volume V with one movable side. The fluid in the box has pressure p . If p is positive, the pressure pushes outwards on the walls. Now apply a force F that moves the movable side inwards, like a piston. Let us suppose that the movement is very small, so that the force F just balances the pressure. Then we can calculate F from the fact that the pressure is the force of the gas per unit area on the wall. If the wall has area A then we simply have $F = pA$. Now, if the wall moves a small distance δx , then the work done by this force is

$$W = F\delta x = pA\delta x.$$

The product $A\delta x$ is the reduction in the *volume* of the box. So the change in volume is $\delta V = -A\delta x$, the minus sign indicating that the volume of the box has been reduced. The result of all this is a very general law about work on gases:

$$\text{work done on a gas to change its volume} = -p\delta V. \quad (19.8)$$

Now, the work done by the external force F is work done against the atoms or molecules of the gas. The pressure is nothing more than the result of untold numbers of collisions between gas particles and the walls. So when the wall moves inwards, it pushes a little on each particle that it encounters, making it rebound from the wall

a little faster than if the wall had not been moving. So the work done by F equals the increase in kinetic energy of these particles, in close analogy with the situation we met when we first introduced the concept of work in Chapter 6. There the work done by the gravitational force increased the kinetic energy of a body orbiting the Sun, as in Equation 6.19 on page 62. In a fluid, however, collisions among the gas particles themselves quickly transfer this energy around the fluid, sharing it roughly equally among all the molecules. The result is that the energy of the fluid has increased by the amount of work done by the external force.

Now, let us look at our box full of cosmological fluid. Its initial energy content was $\rho_\Lambda c^2 V$. During the expansion, the pressure did not change, and the volume changed by V . The work done, and hence the change in the total energy in the fluid, is $-p_\Lambda V$. Since $p_\Lambda = -\rho_\Lambda c^2$, the energy in the box has increased by $\rho_\Lambda c^2 V$, so that the new total energy is $2\rho_\Lambda c^2 V$. But the volume is now $2V$, so the energy density is $\rho_\Lambda c^2$ and the mass density is ρ_Λ , unchanged from before the expansion. This demonstrates that, as the Universe expands, it is consistent with local conservation of energy (i.e. conservation of energy in any small region of the Universe) that the cosmological constant should not change.

The constancy of energy with volume also explains how the mass-energy density of the fluid is independent of the observer. To see this, we suppose that an observer at “rest” measures mass density ρ_Λ , and that, as above, this fluid is in a rectangular box. This time the box does not have a movable wall. Suppose another observer moves at a small speed v along one edge of the box. This observer will notice two things. First, the box is, of course, moving at speed v past him. The fluid in it therefore should have more mass-energy than when it is at rest, because it has the kinetic energy of its motion. However, in this particular case, as we have seen in the text, the inertial mass density of this fluid is zero: it can be accelerated for free, without any energy cost. (The energy of the box, made of ordinary matter, is not of interest to us here.) The second thing the observer will notice is that the box is shorter, because of the Lorentz-Fitzgerald contraction (Chapter 16). Normally this would raise the density of mass-energy in it, but we have seen above that for this kind of fluid, changing the volume of the fluid has no effect on its density. The net result is that the density of this fluid is invariant under a change of observer, just as is the pressure.

So the cosmological constant is quite remarkable: it provides an all-pervading energy density and negative pressure that are the same to all observers, at all places, and at all times in the history of any universe model, even expanding ones.

Exercise 19.2.1: Upper bound on the cosmological constant

The fact that Newtonian gravity describes the orbits of planets in the Solar System very well, using only one parameter (the mass of the Sun) for all planetary orbits, suggests that the cosmological constant must create a smaller mass density than the mean mass of the Solar System out to Pluto’s orbit. (a) Calculate this mean density by dividing the mass of the Sun by the volume of a sphere whose radius is the radius of Pluto’s orbit. (b) From this, calculate the value of the cosmological constant Λ that would give a mass density ρ_Λ of the same value. Use Equation 19.7 on page 255.

that a geometrical theory requires ten equations, in which the ten coefficients are determined by the properties of the source of the gravitational field: energy, pressure, and so on.

Einstein had another problem, though: as we have remarked earlier, the geometry of spacetime does not uniquely determine the values of the coefficients, since we are free to change the coordinates that we use to describe spacetime. Indeed, Einstein wanted to build into his theory this freedom to choose coordinates. But that meant that the equations of the theory could not possibly determine all ten metric coefficients in terms of the sources. If they did, that would amount to determining the coordinates too.

Einstein’s breakthrough came when he found that he could write down ten equations that were not all independent. He could derive some of the equations from the

others, provided the sources of the gravitational field obeyed the laws of conservation of energy and momentum in any locally flat coordinate patch. This was a very significant step. It meant that his geometrical gravity would respect the equivalence principle completely, so that not only would freely-falling particles “feel” no gravitational field, but also freely-falling gases would behave just as if there were no gravitational field. At the same time, Einstein’s equations would only determine six combinations of the metric coefficients, the remaining being determined by coordinate choices.

This requirement, that gravity should be compatible with energy conservation in ordinary matter, almost fully determined the equations of the theory.

The elegance and beauty that mathematicians and physicists find in general relativity comes partly from the way Einstein started with an apparently horrible prospect, namely trying to find ten equations that would work in any coordinate system and that would predict Newtonian gravitational effects when gravity was weak, and yet managed to arrive at a theory that does all this and can be written down in one line.

Here it is:

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = (8\pi G/c^4)T_{\mu\nu}. \quad (19.9)$$

We aren’t going to work with these equations, of course. But it would be a shame to spend a lot of time in this book discussing the theory and repeatedly mentioning the equations, without ever writing them down in their most general form!

The fundamental unknown quantities here are the ten metric coefficients that describe the spacetime-interval, denoted here by $g_{\mu\nu}$, and called the **metric tensor**. The symbol $G_{\mu\nu}$ on the left-hand side is related to the curvature of spacetime, and is constructed from the ten metric coefficients; its name is the **Einstein curvature tensor**. The idea is to solve these equations for the metric coefficients as functions of position in space and time, given (as the source of gravity) the density, pressure, and momentum of any matter fields that are present: the first three of our sources are on page 240. These sources are all part of the object $T_{\mu\nu}$ on the right-hand side, which is called the **stress–energy tensor**. This is the single source to which we referred earlier in the chapter, when we described the symmetry of the principle of general covariance.

The constant Λ is Einstein’s cosmological constant. He placed it on the left-hand side of Equation 19.9, as part of the equation to be solved, rather than on the right, with the sources. We noted why earlier.

What makes the Einstein equations mathematically challenging is not just that they use the language of tensors. More important is that the Einstein curvature tensor can only be constructed from the metric by using calculus. It is a function of the **derivatives** of the metric tensor. It is a non-linear function, so it is in the Einstein tensor that our fourth “source” on page 240 is to be found. The Einstein equations form a set of what mathematicians call **differential equations**, and their complexity is so great that they can be solved by algebraic methods only in special circumstances, such as when one is looking for solutions with a particular symmetry. Full solutions, for example those that represent collisions of black holes, must be solved on supercomputers.

The search for simplicity

The fact that Einstein’s equations can be written in a single line, using only a few symbols, is a reflection of the fact that they are, conceptually, simple equations. The symbols $G_{\mu\nu}$ and $T_{\mu\nu}$ refer to meaningful combinations of mathematical entities

►The word **tensor** used in these names refers to a mathematical object that is a generalization of a **matrix**, or array. The symbols μ and ν are labels (called **indices**) that can together be taken in ten different combinations to make the full set of equations. The word **stress** is a physicists’ word for things like pressure and tension.

This is how the pressure contributes to the creation of gravity, as we have seen it does in general relativity.

In this section: we meet Occam’s razor and show how Einstein used it when devising, and later revising, the field equations.

that were known to physicists and mathematicians even before Einstein; they are not just shorthand for long strings of algebra.

In arriving at his famous equations, Einstein followed a long-cherished principle in science, called **Occam's razor**: “It is vain to do with more what can be done with less.”[†]

Named for the Englishman William of Occam (1300–1349), who was putting into words what had already been practiced by Greek scientists long before, this principle is interpreted by physicists today to mean that, when trying to find a new theory to fit some observed facts, one should always aim for the simplest description.

Inevitably, there will be many theories that might fit the facts, including the trivial one that says, for example, “It is a law of Nature that the Earth should take one year to go around the Sun and it is another law of Nature that Venus should do so in 0.72 years.” Such a “theory” merely re-states observed facts without offering explanations or relations between them, and so is unsatisfactory. Newton’s law of gravity explains these facts and the other planetary periods, plus much more, using only one observed fact, the mass of the Sun. This is an illustration of the simplicity of the theories of physics.

Einstein’s original field equations had no cosmological constant, so they had a simplicity similar to Newton’s gravity, in that they introduced no new measurements or important numbers that are not already present in Newton’s gravity and special relativity. The original equations just use the constants G and c plus the mathematics of spacetime curvature. In a very real sense, Einstein’s theory is the simplest theory that makes Newton’s gravity compatible with special relativity and the other laws of physics.

Later, when Einstein felt compelled by astronomical evidence to introduce the cosmological constant, he retreated a little from the initial simplicity of the theory. But he still used Occam’s razor: he found a way of introducing a cosmological repulsion, or anti-gravity, that did not require any special observer or coordinate system. The very special and peculiar properties that this cosmological fluid possesses allowed Einstein to keep the principle of general covariance and avoid introducing anything new except Λ . This shows that Occam’s razor is not so conservative that it prevents innovation and the modification of old and inadequate theories. Instead it imposes a form of discipline, keeping the innovations as simple as the new facts allow.

General relativity

Our approach in this chapter to the field equations – taking them apart and looking separately at the most important sources of gravity – has given us considerable insight into general relativity, but there are some aspects of the theory that this method does not directly illuminate. To fill in these gaps, here is a partial list to help us get ready for later chapters.

- *Einstein’s equations predict gravitational waves.* This is inevitable in a theory that obeys special relativity and embodies Newtonian gravity. Since no influence, not even a gravitational one, is allowed to travel faster than light, it follows that the changes in a gravitational field that are caused by changes in its source (such as the orbital motion of a pair of binary stars) must travel outward no faster than light. This outward motion of the changes of gravity is

In this section: we step back and look again at the theory from Einstein’s point of view. We mention some predictions, like gravitational waves and cosmology, that we have not looked at so far in this chapter.

[†]*Frustra fit per plura, quod fieri potest per pauciora.*

like a wave moving along the surface of a pond from the point where a stone falls into the water. We call this a gravitational wave. In general relativity, gravitational waves move at exactly the speed of light. Chapter 22 is devoted to gravitational waves.

- *General relativity can deal consistently with cosmology.* The idea that gravity is geometry rather than an extra force has this unexpected and useful side-effect. Newton's rule for computing gravity makes no sense if one tries to apply it to an infinite Universe, where one has to add up the effects of an infinite number of galaxies. General relativity avoids this problem because gravity is just geometry; it does not add up direct long-range forces. As long as the geometry of the Universe is smooth, then gravity evolves with time in a regular way, regardless of how big the Universe is. Moreover, since changes in gravity move at a finite speed (c), very distant parts of the Universe do not affect us. If they are so far away that light could not have reached us since the Big Bang, then they can have no influence on our local geometry. We shall see that, in fact, all we need is our local weak-field equations to compute the geometry of cosmology.
- *The Einstein equations are not the only ones that one might invent.* There are more ways than one to write down a generally covariant set of equations for a curved spacetime that satisfies the equivalence principle. But the success of general relativity in experiments and Solar System observations has shown that any changes to Einstein's equations need to be small. They are most likely to arise from the next item on our list, **quantum gravity**.
- *General relativity is not a quantum theory of gravity.* Planck's constant is conspicuously absent from Einstein's equations. There is therefore no uncertainty principle: all gravitational effects can, at least in principle, be measured with arbitrary accuracy. This can lead to logical contradictions with the rest of physics, if for example one imagines using gravitational means to measure the positions of elementary particles. Since we believe that the Universe is basically quantum in nature, we expect that general relativity will ultimately have to be replaced by a quantum version. It is likely that this will effectively change Einstein's equations by adding correction terms proportional to Planck's constant. We will return to this subject in Chapter 27.

Looking ahead

In this section: we are ready to apply the principles of general relativity to the systems discovered by astronomers.

We have now laid the foundations for the remainder of this book. We have opened the door to the rich and fascinating world of relativistic gravitation. We have had hints before, about black holes and neutron stars, about gravitational collapse and gravitational waves, about inflation and the Big Bang. But now we are in a position to understand these ideas and phenomena in a deeper way. We will start with neutron stars, progress to black holes, look at the new astronomy that gravitational wave detectors will soon open up, learn how the deflection of light is being used to discover dark masses, and then confront the ultimate: cosmology, the Universe as a whole. Almost every proton, neutron, and electron in our bodies has been in existence since about three minutes after the Big Bang. We are ready now to begin to understand the history of the matter of which we are made.

Neutron stars: laboratories of strong gravity

In previous chapters, we have seen how the new ideas in Einstein's gravity make small but striking corrections to the predictions of Newton's gravity, bending light more strongly as it passes the Sun and causing the orbits of planets to precess. Working out these corrections helped to ease us into the theory, to see that relativistic gravity is a natural development from Newtonian gravity. But the real excitement in modern astronomy and theoretical physics is in situations where Newtonian gravity doesn't even come close to being right. The Universe demands that astronomers use general relativity to explain what they see, and the deepest questions of fundamental physics demand that physicists even go beyond general relativity to find their answers. In this chapter we open the door on the richness of modern gravity by studying our first example of really strong gravitational fields: neutron stars.

Neutron stars are effectively giant nuclei, held together by gravity. If Isaac Newton had understood enough nuclear physics, he could have predicted their existence, and he could have given a rough description of them within his theory of gravity. We did this in Investigation 12.6 on page 148. When we look at this below, we will see that such a calculation merely shows us that Newtonian gravity cannot give a particularly accurate description of neutron stars: relativity cannot be ignored or relegated to a small correction.

What Newton also would not have been able to do, even with the best nuclear physics, is to have predicted how *abundant* neutron stars are. Possibly one star in every thousand in our Galaxy is a neutron star. Newton also could never have guessed how spectacularly they show themselves off, as pulsars and intense sources of X-rays. Containing more mass than the Sun, in a region smaller than a large city, a typical neutron star spins on its axis tens of times per *second*, nurtures a magnetic field billions of times stronger than the Earth's, and – with an interior temperature of millions of degrees or more – is the ultimate high-temperature superfluid and superconductor.

Nuclear pudding: the density of a neutron star

In Investigation 12.6 on page 148, we calculated roughly some of the properties of neutron stars from basic quantum theory and Newtonian gravity. Here we take a different point of view, and show that, without knowing much about quantum theory, it is easy to see that a neutron star should have the same density as an ordinary heavy nucleus, like that of uranium. It may seem strange to try to extrapolate from a tiny nucleus to a huge neutron star, but nuclei have one unusual property that makes this possible.

This property is that the nuclei of *all* heavy elements have very similar densities, about $2 \times 10^{17} \text{ kg m}^{-3}$. This is a huge density by ordinary standards, some 2×10^{14} times the density of water. Since almost all the mass of an atom is concentrated in its nucleus, the nuclei occupy a very small part of the volume of an atom, smaller in

In this chapter: we study neutron stars, our first example of strong relativistic gravity. Neutron stars are known to astronomers as pulsars and X-ray sources, and they are at the heart of supernova explosions. They are giant nuclei containing extreme physics, including superstrong magnetic fields, superconductivity, and superfluidity. Neutron stars only exist because of a few coincidences among the strength of the nuclear, electric, and gravitational forces; without these coincidences, life would never have formed on Earth.

>Underlying the text on this page is a sketch of a *pulsar*, which is a spinning magnetic neutron star. The magnetic field lines (arcs) converge on the magnetic poles, which are hot spots, emitting beams of radio, visible, X-ray, and gamma-radiation. The magnetic poles lie near the equator of the spinning star, whose spin axis might point vertically on this page. The poles sweep the sky like a lighthouse, so that if the Earth is in one or both beams, we see the star turn on and off. Figure 20.4 on page 270 shows a photographic record of the light from such a star flashing 30 times a second.

In this section: neutron stars are simply huge nuclei, held together by gravity rather than nuclear forces. Their existence depends on the push-pull nature of the nuclear forces, which stop gravitational collapse when the protons and neutrons get about as close to one another as in a normal nucleus.

size by the cube-root of 10^{14} , or about 5×10^4 : the radius of the nucleus of an atom is roughly 50 000 times smaller than the orbital radii of its electrons. Put graphically, if the nucleus were magnified to the size of an apple, then its electrons would be 1.6 km (1 mile) away! All the space between the nucleus and its electrons is empty.

Now, because nuclei have this same high density no matter how many neutrons and protons go into the nucleus, every **nucleon** (every proton or neutron) occupies the same volume as every other one, and this volume is virtually the same, regardless of whether there are 10 nucleons or 100.

This is a rather unexpected behavior: one would normally expect that the nuclear forces that attract nucleons together would get stronger as more nucleons are added, and the density would increase. This happens when gravity provides the attractive force and the matter is ordinary gas: when more mass is added to a star, its density normally increases. So nuclear forces must be different from gravity somehow, to keep the nuclear density at its special value.

The difference can be understood by analogy: if we fill a box with plastic balls, the density of balls (number per unit volume) does not change with the number of balls we add. Balls keep piling up, but, as long as none of them gets crushed, the density is determined only by the size of each ball. A small box with 10 balls and a large box with 100 balls will have roughly the same density. The reason is that the balls are hard: when they get sufficiently close to one another, they resist being pushed any closer.

The uniform density of nuclei means that the nuclear forces must have a *hard core* of repulsion that keeps nucleons a certain distance apart. When nucleons are further apart than the size of this hard core, the nuclear forces attract them together. This attraction holds nuclei together against the repulsive force of the positive electric charges on all the protons. But the nuclear attraction must change to repulsion when the nucleons get sufficiently close.

The nuclear density quoted above tells us that each nucleon of mass 1.67×10^{-27} kg occupies a mean volume of about $8 \times 10^{-45} \text{ m}^3$, which is the volume of a cube of side $2 \times 10^{-15} \text{ m}$. Now, the nucleons will be separated by the sum of both repulsive hard cores, so the radius of the hard core should be no more than half the side of this cube, 10^{-15} m . The core radius has been measured experimentally to be about $4 \times 10^{-16} \text{ m}$. This is consistent with our estimate: one would expect nucleons to keep a little further apart than their minimum core radius, since in a nucleus they form a quantum Fermi gas (recall Chapter 12) in which the nucleons move around and have a quantum uncertainty in their positions.

Physicists do not clearly understand the forces between nucleons when they are pushed up against this core. Much research in modern nuclear physics is directed at understanding the attractions and repulsions between nucleons at short range, and some of the tools of that research are giant accelerators that can smash heavy nuclei together to form super-dense collections of hundreds of nucleons. But, if we are explaining the density of neutron stars, we can take the basic hard core as a starting point.

We saw in Chapter 12 that when white dwarf cores of giant stars collapse, the high densities force electrons and protons to combine into neutrons. Yet this in itself does not make a neutron star. The object can become a star in equilibrium only if it can support itself against gravity. Because the nuclear forces are attractive until they reach the hard core, this support can happen only if the density is close to the nuclear density.

Here is the argument in detail. Suppose we have a gas of neutrons whose density is much less than nuclear density. Then when neutrons collide with one another, as must happen at random all the time, they will have a tendency to stick together and form large nuclei; the extra pressure of neutrons from outside will make these condensations bigger than ordinary nuclei like uranium. So the gas will be a mixture of free neutrons and big nuclear lumps. Now consider what happens when the gas is compressed. Collisions become more likely, and the result will be that many neutrons will get stuck in lumps and not contribute to the gas pressure: remember from Chapter 7 that the pressure of a gas depends on its temperature and the number of particles in the gas, not on the masses of the particles. If free neutrons are lost to the lumps, then the pressure will not build up fast when the gas is compressed, and it will not be able to hold itself up against gravity. The collapsing core of a giant star in a supernova explosion therefore continues to collapse well after neutrons have been formed.

When the density of the collapsing core reaches nuclear density, the lumps all merge into a smooth “pudding”, and further compression sees a rapid increase in pressure from the hard-core repulsion. Collapse stops, and a neutron star with the density of an ordinary nucleus is formed.

Now we see that the incredibly small size of neutron stars compared to the Sun is not so hard to explain. In normal matter, nuclei are separated by the huge distances occupied by the intervening electrons. In a neutron star, Nature has simply managed to remove all that wasted space, and put all the nuclei right up against one another.

It takes a whole star to do the work of 100 neutrons

This argument tells us what the density of a neutron star should be, but it does not tell us about the mass. It does not tell us whether this phenomenon should occur with stars or with basketballs. Why are there neutron stars, and not neutron basketballs? Or neutron galaxies?

The answer, of course, is gravity. To see why, let us imagine trying to make a neutron basketball. If one takes a heavy nucleus, say of uranium, and tries to build it up into a neutron basketball by adding one nucleon at a time, something goes wrong: as soon as a nucleon is added, the nucleus spits it out again, or worse still the whole nucleus divides in half. This is *radioactivity*.

Heavy nuclei decay through radioactivity because they are unstable. This happens basically because of the second feature of the nuclear force that did not come into our previous discussion but which is obvious if we look at everyday life from the right point of view: even the attractive part of the nuclear force is of very *short range*. We can see that it must be short range, since essentially all the properties of ordinary materials can be explained by using just the electric and magnetic forces that electrons and nuclei exert on one another through their electric charges. The nuclear forces are intrinsically strong, since they can hold all the protons in a nucleus together, despite their mutual electric repulsion. But they do not extend very far from the nucleus, since they do not influence chemistry. Unlike gravity and the electrostatic force, which fall off as $1/r^2$ as one goes away from the source, the nuclear force must fall off much more rapidly as one leaves the nucleus.

This means that, as one adds nucleons to a nucleus, there will come a point where nucleons on one side of a nucleus no longer feel the attraction exerted by those on the other side. Protons still feel the electrostatic repulsion of other protons, however, so if one adds protons to a sufficiently large nucleus, they will simply be pushed out

► Readers who have read Investigation 8.8 on page 101 will have already gone through this argument about pressure in detail.

In this section: the neutrons in a neutron star are held together by gravity. We show here that gravity is only strong enough to replace the binding forces that hold nuclei together when there are as many nucleons as in a typical star. This coincidence is one of the deep mysteries of nature, because without neutron stars there would be no life on Earth.

again: the new proton feels a nuclear attraction from only a few nucleons, but a repulsion from all the existing protons.

If one adds neutrons, one avoids this repulsion, but one still runs into a problem: the Pauli exclusion principle (Chapter 12). As one adds more and more nucleons of either kind, the new ones cannot have the same low kinetic energies of the existing ones, since the existing ones have filled up all the low-energy quantum states. So new nucleons must have higher energies, and at some point these energies will be enough to escape from the attraction exerted by the nearby nucleons. At this point, the nucleus will accept no new nucleons. This happens at roughly an atomic number of about 210: nuclei with more than 210 neutrons and protons in total tend to be unstable. This is about the location of lead in the periodic table.

So we are frustrated in our attempts to build a nucleus with the mass of a basketball by the short range of the nuclear forces. To hold a bigger nucleus together requires a long-range force, and the only candidate is gravity. Electric forces won't do, since like charges repel, and an equal mixture of positive and negative charges will exert no net long-range force. So only gravity can stabilize nuclei bigger than lead.

Yet gravity is a weak force, and the attraction it exerts within an ordinary nucleus is tiny compared to the other forces. Gravity can only provide the glue to hold together a large nucleus if the self-gravitational force of the nucleus is comparable to the nuclear forces between nuclei. This is going to require a large amount of mass.

We can in fact compute just how much mass is required by a relatively simple argument. It is observed experimentally that the "escape energy" of a nucleon from a nucleus is about 8 MeV, the same for most nuclei. This is the energy that has to be supplied to a nucleon to get it away from the nucleus, and nuclei become unstable when the exclusion principle forces new nucleons to have this energy inside the nucleus. A nucleon that has a kinetic energy of 8 MeV inside a nucleus has just enough speed to escape. We show in Investigation 20.1 that this escape speed is about 13% of the speed of light.

Now, gravity can prevent this escape if it raises the escape speed: if the escape speed from a large clump of neutrons exceeds this value, then the clump will be one big stable self-gravitating nucleus: it will be a neutron star. In Investigation 20.1 we show that a star with the density of a nucleus has an escape speed exceeding the nuclear escape speed if the mass of the star exceeds roughly $0.02M_{\odot}$.

An object with more than 2% of the mass of the Sun and the density of a nucleus has strong enough gravity to keep the nucleons bound together.
This is our estimate of the minimum mass of a neutron star.

So there are no neutron basketballs.

Despite the simplicity of our argument, our estimated minimum mass is very close to the value of $0.1M_{\odot}$ that full calculations give in general relativity, using more sophisticated nuclear physics. Considering that we have bridged a gap of a factor of 10^{53} from a nucleus of mass, say, 10^{-25} kg to the mass of a star, to have come within a factor of five of the right result is close indeed!

What about the *maximum* mass of a neutron star? As for white dwarfs, the maximum mass for neutron stars is set by the balance between the inward pull of gravity and the amount of pressure the nuclear matter can sustain. In Investigation 12.6 on page 148 we calculated the maximum mass of neutron stars in the same way as we calculated the Chandrasekhar mass for white dwarfs, and we obtained the result that the maximum mass should be about five or six solar masses.

► Lead is an abundant mineral on Earth because it has been produced by the radioactive decay of heavier nuclei over the ages.

► Recall that the symbol "MeV" represents a million electron volts, which is 10^6 eV = 1.6×10^{-13} J.

► Thoughtful readers will realize that our argument here is certainly an oversimplification, since nuclei are not electrically neutral, and the electric repulsion of the protons must affect their structure and in particular the escape energy.

Neutron stars are neutral, so their escape energy will depend only on the nuclear forces. However, since the neutrons in a nucleus are not affected by the nuclear force, and still they have escape energies comparable to those of the protons, the argument here should be accurate to within an order of magnitude.

► Notice that our way of calculating the mass of the neutron star from the binding energy of a nucleus is essentially the same argument as we used in Investigation 8.3 on page 91 to calculate the minimum mass of an object in the Solar System that is round, from the binding energy of silicon dioxide.

Investigation 20.1. Minimum mass of a neutron star: no basketballs

Experiments show that the energy required to remove a nucleon from an ordinary stable nucleus is about 8 MeV. From this energy it is possible to deduce an "escape speed" for a nucleon from the formula

$$K = \frac{1}{2}mv^2,$$

where K is the escape energy. Some arithmetic gives, using the mass of the proton for m (see Appendix A),

$$v_{\text{escape}} = 4 \times 10^7 \text{ m s}^{-1},$$

or 13% of the speed of light.

Now, we want the escape speed from the neutron star to exceed this. This speed is (in Newtonian gravity) $(2GM/R)^{1/2}$, where M is the mass of the star and R its radius. The star must have nuclear density ρ_{nuc} , which means that we can deduce its radius from its mass. Writing down the equation for the average density,

$$\rho = \frac{M}{\frac{4}{3}\pi R^3},$$

Exercise 20.1.1: How big is the nuclear hard core?

Use the mass $1.67 \times 10^{-27} \text{ kg}$ of a nucleon and the density $2 \times 10^{17} \text{ kg m}^{-3}$ to calculate the volume occupied by each nucleon in a nucleus. If the nuclei are contained in cubical boxes, how big is each box? What is the size of the hard core, the irreducible radius of a nucleon?

Exercise 20.1.2: Calculating the minimum neutron star mass

Solve Equation 20.1 for M and use the value of ρ_{nuc} in the previous exercise to verify the minimum mass in Equation 20.2.

Exercise 20.1.3: What does a neutron star look like?

Taking the mass of a neutron star to be $1M_{\odot}$, what is its radius? What is the escape speed of a projectile leaving its surface? What is the speed with which a projectile falling from rest far away reaches the surface? What fraction of the rest-mass of such a projectile is its kinetic energy when it arrives at the surface? What is the orbital speed of a particle in a circular orbit just above the surface of the star? What is its orbital period? Do all calculations using Newtonian gravity, even though the speeds are relativistic.

Exercise 20.1.4: Thermal effects in neutron stars

If the binding energy of a nucleon is 8 MeV, what temperature would the star have to have in order to boil off a nucleon? Since the pressure support for the star comes from the hard-core repulsion and not from random thermal motions of the star, it is possible for stars to cool off after formation without changing their properties. Give an argument that a star is "cold" (thermal effects are unimportant for its structure) if its temperature is smaller than the one you have just calculated. Assume the star has a temperature of 10^6 K . What is its black-body luminosity? (See Equation 10.3 on page 116.) What is the wavelength at which it is brightest? (See Equation 10.9 on page 117.)

However, this is too simple an estimate, since Newtonian gravity is just not accurate enough for such compact stars. One needs to use general relativity to calculate their structure. We will do this in Investigation 20.3 on page 280, but for now we only point out that the effect of using relativity is to lower the maximum mass to somewhere between two and possibly three solar masses. Its actual value is not known: uncertainties in nuclear physics prevent reliable calculations.

We have learned that neutron stars can only exist in a rather restricted range of masses, between perhaps 0.1 and two solar masses. In fact, their lower limit in practice will normally be much larger, since a collapsing star will stop at the white dwarf stage if its mass is less than the Chandrasekhar mass. Neutron stars should only form if their masses are somewhat larger than $1M_{\odot}$. A collapsing star above the maximum mass will continue to collapse, and will form a black hole.

It is also interesting to ask what happens if we have a neutron star that subsequently gains or loses mass. If it gains mass, perhaps from a companion in a binary system, then it can be tipped over the maximum and it will collapse to a black hole. If it loses mass, again to a companion in a neutron star binary (see below), then when it reaches the minimum mass it will no longer be bound together and will undergo a catastrophic nuclear disintegration: it will explode.

The most remarkable and fortunate coincidence about these masses is that the maximum mass is larger than the Chandrasekhar mass. This coincidence allows

and solving for R , we find

$$R = \left(\frac{M}{\frac{4}{3}\pi\rho} \right)^{1/3}.$$

This gives a gravitational escape speed for a "nucleus" of mass M :

$$v_{\text{escape}} = 1.8G^{1/2}M^{1/3}\rho_{\text{nuc}}^{1/6}. \quad (20.1)$$

Using the value of the nuclear escape speed for v_{escape} here gives a minimum value for M , which is

$$M_{\text{min}} = 4 \times 10^{28} \text{ kg} = 0.02M_{\odot}. \quad (20.2)$$

neutron stars to form in the first place. The maximum mass is a property of nuclear physics and general relativity. The Chandrasekhar mass depends on Newtonian gravity and atomic physics. We could imagine a Universe in which the nuclear repulsive core was smaller, so that neutron stars were denser and the effects of general relativity correspondingly greater, leading to a maximum mass smaller than the Chandrasekhar mass, which is unaffected by the nuclear hard core. In such a Universe, collapsing stars bigger than white dwarfs would form black holes directly. And in such a Universe, people would not exist.

The reason is that the nuclear hard-core repulsion plays a key role in the chain of events that leads to life on Earth. We have seen that the elements of which we are made were formed in stars, and that the heavier elements are spread into interstellar clouds by supernovae. Our Sun and Earth formed from clouds seeded with oxygen, silicon, and many other elements essential for life, by a long-ago supernova. But that supernova could not have happened if neutron stars could not form. If the collapsing core of the giant star that became the supernova could simply have continued to collapse to a black hole, then there would have been no “bounce”, no shock wave to blow off the envelope of the giant star. Instead, all the gas in the giant star would have fallen into the black hole. The vital elements carried by the supernova gases would never have left the star and found their way into our Solar System.

We owe our existence to the existence of neutron stars, and in particular to the neutron star that formed in that long-ago supernova. We must be thankful that Nature has arranged for the Chandrasekhar mass to be smaller than the maximum mass of neutron stars.

What would a neutron star look like?

In this section: from the properties of neutron stars we have already calculated, it is possible to make predictions about them. They are clearly dense and hot. They should emit X-rays and they should spin very fast. They could have strong magnetic fields.

Let us ask a few questions about the typical properties of a neutron star, just using the numbers we have obtained so far and assuming we can use Newtonian gravity.

Let us suppose the star has a mass of $1M_{\odot}$. This can't be far wrong, since a star with a mass less than the Chandrasekhar mass (see Chapter 12) will support itself at the density of a white dwarf and not collapse to a neutron star. So by taking a mass of $1M_{\odot}$ we are probably underestimating a little, but it will suffice to give us an idea of what the star will be like. In Exercise 20.1.3 on the preceding page you have the opportunity to do the calculations leading to the numbers below.

With a density of $2 \times 10^{17} \text{ kg m}^{-3}$, the star's radius will be about 13 km, or 8 miles: it would just cover Manhattan Island. The escape speed for a particle at its surface is $1.4 \times 10^8 \text{ m s}^{-1}$, or about half the speed of light. The orbital speed of a satellite at its surface is about 10^8 m s^{-1} , one-third of the speed of light. The period of such an orbit is 0.8 ms. This also sets the maximum spin rate of a neutron star: it could in principle rotate about 1000 times per second without flying apart.

The compactness of the star tells us also that the gravitational redshift of light from its surface will be significant. By the equivalence principle, an observer falling freely from far away will see no change in the frequency of light. Such an observer reaches the surface with the escape speed, $c/2$. The redshift seen by observers that remain at rest with respect to the star, then, is the same as that seen by an observer who is receding from a source of light at this speed. This will produce a lengthening of the wavelength of the light by at least a factor of two. If astronomers could see spectral lines in the radiation from a neutron star, they should be strongly redshifted.

These numbers can only be approximately correct of real neutron stars, since they tell us that typical speeds associated with the star are good fractions of the

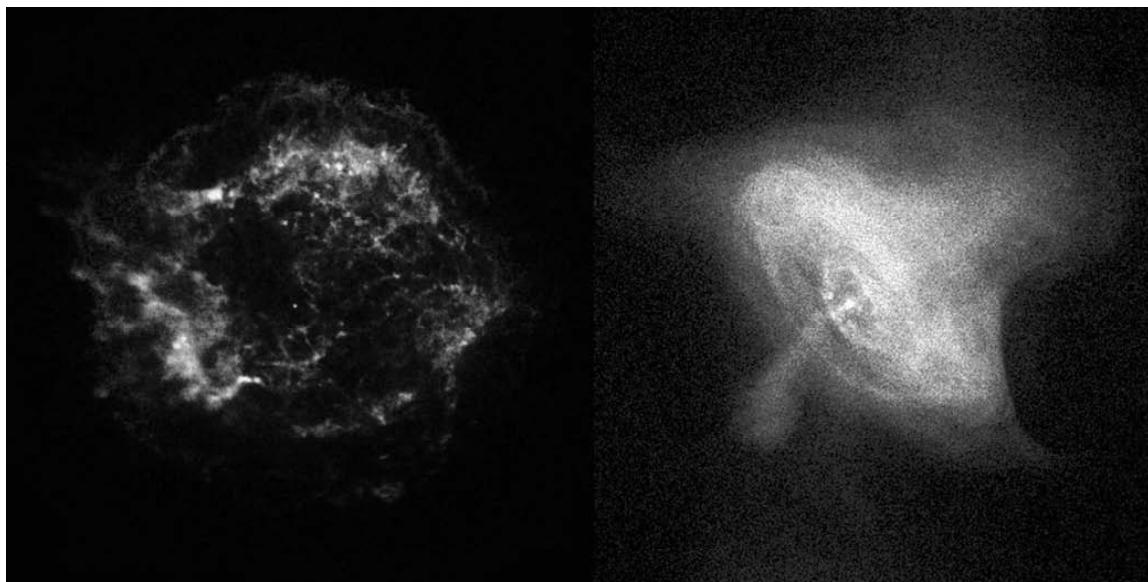


Figure 20.1. The Chandra X-ray satellite has excellent sensitivity and imaging capability. It took these two photos of supernova remnants. The one on the left is called Cassiopeia A, which is a supernova that exploded about 300 years ago. The X-rays reveal the expanding cloud of gas and, for the first time, a bright spot at the center. Further studies will reveal whether this spot is thermal radiation from a neutron star or from a disk of gas falling into a black hole; either kind of object might have been created. The photo on the right is of the central part of the Crab Nebula; an optical photo of the entire nebula is shown in Figure 20.4 on page 270. This nebula was formed by a supernova explosion about 1000 years ago, and it contains a neutron star. The photo reveals a complicated cloud of gas around the neutron star, and very interestingly a jet of gas shooting out from the neutron star. This is a miniature form of the phenomenon seen in quasars. The Chandra satellite is named after the astrophysicist S Chandrasekhar, whose work was introduced in Chapter 12. Image courtesy of NASA/CXC/SAO.

speed of light. This means that Newtonian theory is suspect, and we really have to use general relativity to get reliable numbers. We shall do that below. But first, let us consider what even these approximate numbers tell us about where we might expect neutron stars to be found and what they might look like.

One thing is clear; the ordinary thermal radiation from neutron stars should not be visible on ordinary photographic plates. Suppose a neutron star has a very high surface temperature, say as much as 10^6 K. Then the black-body luminosity (Equation 10.3 on page 116) is about one-third of the luminosity of the Sun. (See Exercise 20.1.4 on page 265.) Most of this energy comes out, however, near wavelengths of 3×10^{-9} m (Equation 10.9 on page 117), which is in the X-ray band of the spectrum. Only a tiny fraction emerges in the visible region, so we do not expect to see this black-body radiation in photographs, but we should hope to find it with X-ray telescopes. X-rays are the ideal means for studying neutron stars; see Figure 20.1. There are also spectral lines in the X-ray band, for example from ions of iron that have been stripped (by the high temperatures) of all but one electron. One might see these lines strongly redshifted.

Observable property 1: neutron stars should emit thermal X-rays with redshifted spectral lines. The redshift is a *diagnostic*, separating neutron stars and black holes from other possible sources of X-rays.

Another observable property of neutron stars is their short time-scales: with their small size and large velocities, any dynamical process will happen very quickly. If astronomical phenomena are found that involve changes on millisecond time-scales, then one should consider whether a neutron star might be responsible. In

fact, one should not be surprised to find a neutron star that is rotating rapidly, because any rotational speed of the original collapsing star should increase during collapse, just as an ice-skater spins faster when she pulls her arms in. Any spin rate, right up to the breakup speed, would in principle be possible.

Observable property 2: neutron stars can exhibit variability on millisecond time-scales, either from pulsations or from rapid rotation. This is also a diagnostic feature, since only neutron stars or black holes can be compact enough to allow such rapid changes.

A third feature we might predict about neutron stars is that they could have strong magnetic fields. This is because, according to the laws of electromagnetism, when an electrical conductor changes size, the magnetic field it is carrying will change in proportion to the inverse square of the size of the conductor. This is the same proportionality as in the law of conservation of angular momentum. By the same calculation as we did above, a neutron star could have a magnetic field larger than that of its progenitor by 5×10^9 . Ordinary stars have fields of a few gauss, so we might expect to find magnetic fields of billions of gauss on neutron stars.

Magnetic fields in astronomy are usually associated with radio emission: strong radio sources tend to have strong magnetic fields. The fields accelerate free electrons, and when they accelerate they give off electromagnetic radiation. So one might expect that any compact source of unusual radio radiation might be associated with a neutron star.

Observable property 3: neutron stars might have strong magnetic fields, and these could make them strong radio sources. The existence of radio emission is not unique to neutron stars, but if the radio emission indicates a very strong magnetic field or exhibits very rapid time-variability, then this would also be diagnostic of a neutron star.

► Zwicky, introduced in Figure 14.8 on page 171, was led to the idea of neutron stars by his study of supernovae, which he was the first to identify as a special class of phenomena. It is remarkable that, only a couple of years after the discovery of the neutron, Zwicky was prepared to postulate whole stars made of neutrons! See Figure 20.2 and Figure 20.3 for brief introductions to Landau and Oppenheimer, respectively.

In this section: here we follow up our predictions and suggest how one might design observations to find neutron stars.

We have identified three observable properties of neutron stars that could help in finding them. This list could in principle have been made at any time since neutron stars were first predicted, independently in the 1930s by Zwicky (whom we met in Chapter 14) and the Russian physicist Lev Landau (1908–1968). The American physicist J Robert Oppenheimer (1904–1967), inspired by the work of Landau, and working with graduate students, showed convincingly within general relativity that the formation of neutron stars by gravitational collapse was possible and indeed in some circumstances inevitable. Thus many parts of our discussion were understood by the 1950s, and yet they were ignored by most astronomers. Neutron stars still proved elusive to observe, partly because no-one was looking, and partly because the technology available to astronomers was not what was needed.

Where should astronomers look for neutron stars?

If enough astronomers in the 1950s had taken the idea of neutron stars seriously, where might they have looked? Where might they look today? Since we expect neutron stars to be formed in supernova explosions of Type II (which are triggered by core collapse), the first place to look for them is in the clouds of gas that mark the supernova remnants, as in Figure 20.1 on the previous page and Figure 20.4 on page 270. These often – but not always – contain observable neutron stars.

Supernova remnants fade away after a few tens of thousands of years, while neutron stars have been produced in our Galaxy for billions of years. Therefore, most neutron stars should be scattered randomly around the Galaxy. An isolated, old neutron star would be very difficult to observe, being too cool even to give off much X-radiation. So the other place to look for neutron stars would be in binary

systems, where any interactions between a neutron star and its companion might reveal the neutron star.

The problem with binaries is that the supernova explosion that produces the neutron star is likely to disrupt the binary system. This is not caused by the exploding gases themselves: they flow around the companion star and give it hardly any push. But the gases carry away mass, and this reduces the gravitational attraction between the companion and the neutron star that is left behind.

If the companion is a star of small mass, then it is easy to see what happens. The escape speed of the companion from its orbital position is only $\sqrt{2}$ times larger than the circular orbital speed, so if the supernova reduces the central mass enough to reduce the escape speed by the same factor, then the companion will find itself moving with enough speed to escape from the neutron star, and the binary will fly apart. Now, the orbital and escape speeds depend on the square-root of the central mass, so we conclude the following.

If the supernova expels more than half of the mass of the star, then a low-mass binary companion will be expelled and the binary will not survive.

Most supernovae would be expected to expel far more than half of the original mass. If the system does survive, the excess speed of the companion will turn the initially circular orbits into elongated ellipses.

On the other hand, if the companion has much more mass than the original supernova star, then the binary will survive. To see why, turn the argument around and regard the supernova star as the small mass orbiting a large mass. Then if the small mass splits into any number of pieces, each will orbit the companion in the same orbit as the original. If then most of these pieces are sent away by the explosion, any piece left behind will not be affected: its orbit is determined by the companion, not by the other pieces.

So the fate of a binary after a supernova depends very much on the ratio of the masses of the two stars: if the supernova occurs in the less massive of the two, it may remain as a binary, but otherwise probably not.

And even in those binary systems that survive, one would expect the initially circular orbits to have become elliptical. Since the statistics of binaries are poorly known, and the evolution of the stars in them is even less well-understood (particularly in terms of the all-important question of how much mass is transferred from one star to another, or is lost from the system entirely during the evolution of the stars), it is difficult to make reliable predictions at present of how many neutron stars should be in binaries.

One conclusion is safe to draw, however. Since most stars start out in binaries, most neutron stars will have been formed in binaries. Those which then leave the binary do so with a large speed, the escape speed from the orbit. This could be anything up to the escape speed from the surface of a star like the Sun, if the binary orbit had been small enough. This is some 600 km s^{-1} . Observations of pulsars (which are described in the next section) show that they have even higher velocities than this argument would suggest. Some are believed to be traveling faster than 1600 km s^{-1} , and their mean speed is around 400 km s^{-1} . These speeds are actually too high to be explained by orbital breakup, since most binary orbits are fairly widely separated and hence have fairly low velocities. The conclusion that astronomers have drawn is that neutron stars are given a "kick" when they are born,

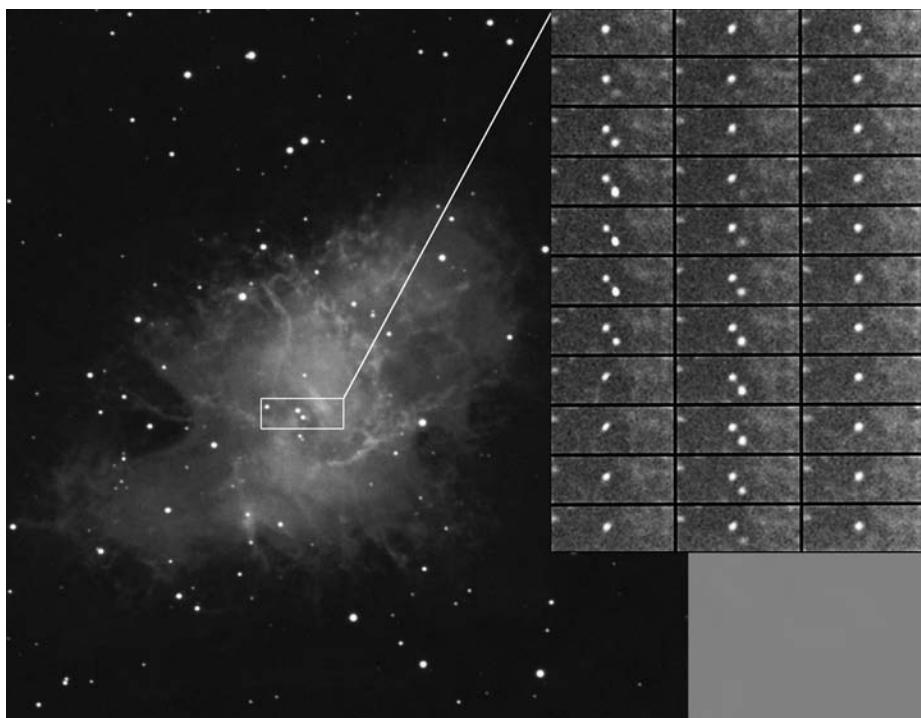


Figure 20.2. Lev Landau was the giant of Soviet physics from the 1930s onwards. His insistence on the highest standard of mathematical ability and his physical insight were legendary. His books and pedagogical legacy are still influential around the world today. Photo courtesy AIP Emilio Segrè Archive.



Figure 20.3. J Robert Oppenheimer's research in theoretical physics and relativity was interrupted by the Second World War, during which he led the American team that developed the fission bomb. After the war he worked to avoid an arms race between the USA and the Soviet Union, and he opposed the development of the fusion bomb. For these stands he was excluded from further advising the government of the USA, and he was treated like a traitor by some politicians and parts of the popular press. The photograph was taken around 1944, when he was leading the bomb project. Image courtesy U. S. National Archives, ARC picture 29-1233a.

Figure 20.4. This composite shows that the Crab Nebula (Figure 12.8 on page 150) contains a pulsar. The series of images on the right show the central two stars, one of which is the pulsar and the other of which is a background star that happens to be in the image. The images were produced digitally by recording only the light arriving at the telescope during a particular millisecond of the pulsar's 33 ms period, and adding up many such images over a long exposure time, all from the same phase of the pulsar's period. Thus, the first image is the brightness during the first millisecond of every period, the second image (below the first) is the brightness during the second millisecond, and so on. The background star remains constant while the pulsar turns on and off twice, which occurs as the two pulsar beams sweep past the Earth. (See the drawing underneath the text on the first page of this chapter.) Most pulsars are not visible in optical photographs, but their radio emission and sometimes their X-ray emission behaves in a similar way. The images were made from a two-hour exposure taken at NOAO Kitt Peak in Arizona in October 1989. Image courtesy N A Sharp/AURA/NOAO/NSF.



averaging around 400 km s^{-1} . This presumably has to do with the turbulent hydrodynamics in the gravitational collapse that forms the star, but there is no widely accepted model of such kicks yet. As we remarked above, kicks could also change the spin of the star.

The space velocity of these pulsars is large. The speed of the Sun in its orbit around the center of the Galaxy is only about 200 km s^{-1} , so some neutron stars will have enough speed to escape from the Galaxy entirely. Others should form a population of neutron stars of high speed, distributed broadly around the Galaxy, not confined to the disk of the Galaxy (the Milky Way) where they were produced.

Although we have not reached firm conclusions about where to find neutron stars, we can be comforted by one further conclusion: there must be a lot of them out there. This conclusion comes from some elementary reasoning. Supernova explosions of Type II are rare events, but from the statistics of supernovae in our and external galaxies, it seems that they have been occurring on average about once every 30–100 years in our Galaxy. Over the lifetime of the Galaxy, 10^{10} years, there have been more than 10^8 supernovas. A good fraction of them must have produced neutron stars, so the total number produced must be around 10^8 . Given that the Galaxy has 10^{11} stars, it follows that one in a thousand stars may be a neutron star. There could be even more, if we have underestimated the supernova rate (hidden supernovae) or if neutron stars can be formed in other ways that are not so spectacular.

Pulsars: neutron stars that advertise themselves

A major reason why neutron stars were not discovered earlier is that astronomical technology was not suited to discovering them. For example, the first astronomical X-ray source to be identified (other than the Sun) was the object Sco X-1, seen in a rocket-borne X-ray experiment in 1962. It is now known to be a neutron star, but the first observations did not have the sensitivity to pick out any of the diagnostic features of a neutron star, such as rapid variability or redshifted spectral lines. Optical astronomers had little chance of identifying a neutron star, because even if it emitted enough light, the long exposures required for photographic plates would have prevented them from seeing any rapid variability. Indeed, the neutron star in the Crab Nebula was known for a long time to be a candidate for the object left behind by the explosion, since its spectrum did not resemble that of an ordinary star. But there was not enough information in the optical spectrum for a confident identification. Radio astronomers had, by our earlier discussion, the best chance, but for many years their telescopes were not designed to pick up rapid variability.

The absence of a positive identification of a neutron star, and the extreme properties expected of them, led most working astronomers to regard them as a theoretician's fantasy, if they thought about them at all! It was therefore a complete surprise when they turned up as *pulsars*.

The story of the discovery of pulsars is well worth remembering, for it shows that, despite scientists' attempts to pursue scientific research in a planned and orderly fashion, some of the most important discoveries arrive in ways that are virtually impossible to predict. Astronomers at Cambridge University in England, led by Anthony Hewish (b. 1924), had constructed a special telescope to look for rapid variations in the radio waves arriving at the Earth from distant radio sources. This was not because they were looking for neutron stars. Rather, the radio sources themselves were expected to be fairly constant, and the expected variations were produced by irregularities in the interstellar plasma and the solar wind that the radio waves pass through before reaching the Earth. The instrument was designed to give information about these plasmas by detecting fluctuations on time-scales shorter than a second. With hindsight, we can see that Hewish had built the first radio instrument capable of identifying neutron stars, but no-one realized this until later.

Most of the observations fit the expected pattern of irregular fluctuations, but a graduate student, S Jocelyn Bell (b. 1943), noticed fluctuations that seemed to be periodic, with a period of about 1 s. After checking that nothing was wrong with the telescope, Bell convinced Hewish and her other colleagues that the radiation was coming from an astronomical source, and in 1967 the result was announced. Soon many other such sources were found, with different periods. The flashing sources were named *pulsars*.

Although our discussion in the previous section makes neutron stars an obvious candidate for pulsars, astronomers at first had to consider many alternatives, such as white dwarfs oscillating in their fundamental mode of radial vibration (as we discussed for the Sun in Chapter 8).

But one property of pulsars proved decisive: the pulsations kept time with remarkable stability over many years. The only motion in astronomy that can keep such good time is rotation: pulsars had to be associated with rotating stars.

The period of rotation then points to neutron stars. No star can rotate faster than the orbital period of a satellite at its surface, and this depends just on the average

In this section: when neutron stars were actually first observed, it was through their strong magnetic fields and rapid spin. With magnetic fields 10^{12} times stronger than the Earth's, spinning many times per second, and immensely strong gravitational fields, pulsars are laboratories of extreme physics.

▷The pulsar phenomenon was so unexpected and puzzling that some astronomers seriously wondered at first if the signals were messages from intelligent beings far away across the Galaxy. But the regularity of the pulses, the power required to produce them, and the discovery of several other pulsars very quickly led to the conclusion that the phenomenon was a natural one. Nevertheless, during the first couple of years astronomers frequently – and only half-jokingly – used the acronym LGM for these objects, an abbreviation for Little Green Men!

density of the star. A period of 1 s requires a density of $6 \times 10^{11} \text{ kg m}^{-3}$. This would just barely allow a model based on very rapidly rotating white dwarfs, but when pulsars were discovered with periods as small as 30 ms, the density limit went up to $6 \times 10^{14} \text{ kg m}^{-3}$. Only neutron stars can reach these densities. Astronomers accepted that they had finally discovered rotating neutron stars.

Rotation provides the “clock” that keeps the pulsar ticking regularly, but what provides the pulses of radio waves? The simplest answer is that a pulsar is like a lighthouse: there is a beam shining in a single direction, which is swept past our telescopes by the rotation of the pulsar. Every time it passes us, we see a pulse. The beam emits continuously, but we see it only intermittently. Such a beam can in principle be created by a strong magnetic field, which we saw was a feature we could expect of pulsars. If the magnetic poles of the pulsar’s field are not near the rotational poles, as they are on the Earth, but instead lie in the pulsar’s rotational equator, then what we could be seeing is a view down onto the pulsar’s magnetic pole every rotation.

In this picture, we would expect a second pulse as the other magnetic pole passes our view, and indeed we do see this in many cases. A good example is the Crab pulsar, shown in Figure 20.4 on page 270. However, if the pole lies somewhere between the rotational axis and the rotational equator, then we would see only one pole and not the other. Moreover, for every pulsar we see, there should be many more whose beams never pass over us, and which we therefore do not see.

Although neutron stars were expected by physicists, the idea that they would send out flashing beams to tell us where they are had never been dreamed of! For the discovery of pulsars, Hewish shared the 1974 Nobel Prize for Physics. Many scientists felt that Bell should also have had a share. We may never know why the Nobel committee neglected the key contribution of this graduate student to the project. The Nobel committee seems to have been more sensitive to the contribution of a graduate student when, in 1993, they awarded the prize for the second time to pulsar astronomers for the discovery of the binary pulsar system PSR1913+16. This story will be told in Chapter 22.

The mystery of the way pulsars emit radiation

In this section: the details of how a spinning magnetic neutron star emits radiation in its beams are still unknown. Somehow the magnetic field creates beams of radiation streaming out from the magnetic poles.

Astronomers are in essentially universal agreement about the association between neutron stars and pulsars, and they agree that the beam is related to a magnetic field that is not aligned with the rotation axis. But astronomers agree about little else. In particular, the way in which the magnetic field produces the radiation in its beams is not at all understood.

It may be related to the terrestrial phenomenon of the aurora, in which charged particles from the Sun move along the Earth’s magnetic field toward the magnetic poles, sometimes creating beautiful displays of light as they reach the Earth. In the case of pulsars, the strong magnetic fields (apparently of order 10^{12} G) are able to pull charges off the surface of the star and send them along the field to the poles, so the emission phenomenon is self-feeding.

Pulsars emit more than just radio waves. The beam can contain visible light and X-rays too. The Crab pulsar emits both pulsed X-rays (Figure 20.1 on page 267) and pulsed light (Figure 20.4 on page 270). These images dramatically illustrate just how unusual pulsars are.

The fact that astronomers see optical light from the Crab may seem to contradict our earlier estimate that the thermal radiation from a neutron star would be too weak to be visible. There is no contradiction, because this light is not produced by black-body radiation. It is produced by energetic particles moving at speeds near the speed of light near the magnetic poles of the star. In fact, the absence of an image of

the Crab pulsar during times when the pulse is not arriving is proof that the thermal radiation from this star is very weak.

The rotation rate of pulsars and how it changes

The visibility of the Crab pulsar in so many wavelengths does tell us that there is an enormous amount of energy available for producing this “non-thermal” radiation. Where does this energy come from?

In the end, regardless of the detailed mechanism by which the radiation is produced, the main source of the radiated energy is probably the rotation of the star itself. So as the various forms of radiation carry energy away, the star must slow down. This is in fact seen: most pulsars that have been observed accurately have been observed to change their pulse period slightly over a number of years. The slowing down is very slight, sometimes gaining less than 10^{-15} of a period in each period. Such small changes are measurable only because the pulses themselves are so regular.

Small as the slowing down is, the energy lost by the pulsar is enormous.

In Investigation 20.2 on page 275 we show that the Crab loses about 1.3×10^{30} J of rotational kinetic energy in each period of rotation. This amounts to an energy loss rate about 10^4 times the luminosity of the Sun!

It is worth reminding ourselves where a pulsar’s energy of rotation came from in the first place. The pulsar formed in collapse and, as we pointed out earlier, collapse leads in a natural way to a rapid spin. The energy that goes into the spin comes from the gravitational potential energy released when the star’s core collapsed, so it is gravity that originally supplied the store of energy for this powerhouse.

We should also note that the Crab pulsar is a *slow* rotator: even though it spins 30 times a second, its gravity is strong enough to hold it together even if it spun 1000 times a second. In particular, its shape should be nearly spherical: it won’t bulge out much at the rotational equator.

More than 1000 pulsars are now known, and they have a wide range of periods, down to 1.6 ms, or 600 rotations per second! For longer periods, longer than about 20 ms, the emission seems to be weaker in pulsars with longer periods. It seems that whatever mechanism produces the pulses gets turned off once the star spins down to a certain rate. This accounts for why there are so few observed pulsars when we expect there to be so many neutron stars. Most neutron stars in the Galaxy may well be old, dead pulsars, still rotating at a modest rate but no longer pulsing.

Where does the energy lost from the spin of the pulsar go? Most of it is not going into the pulsed radiation that is seen by radio and optical telescopes: the Crab pulsar emits about 10^{24} J per rotation in each of these, only one millionth of its total rotational energy loss. Most of the energy probably goes into the acceleration of high-energy particles, some of which contribute to the pulses, and into low-frequency electromagnetic waves generated by its rotating magnetic field. Any rotating magnet will emit electromagnetic radiation with a frequency equal to the frequency of rotation of the magnet. For the Crab pulsar, this is 30 Hz.

Waves at such a low frequency get strongly absorbed by the thin plasma that surrounds the star, the remnant of the gas ejected from the supernova star that formed the Crab. So astronomers cannot hope to detect it on Earth. This is a pity, because if this radiation were detected, it would directly measure the magnetic field of the Crab and determine if this radiation really accounts for the energy loss. It is

In this section: pulsars lose much more energy than they put into their beamed radiation, and this energy lights up the gas around them. By losing energy, they slow down, and the rate of slowing allows an estimate of their ages. Most are only a few million years old.

possible that pulsars lose rotational energy in other ways, for example in gravitational radiation, to which we will return in Chapter 22.

If we do assume that the radiation of low-frequency radiation from the spinning magnetic field accounts for the slowdown, then it is possible to calculate the magnetic fields of pulsars. Most pulsars, like the Crab, have field strengths of the order of 10^{12} G. (Remember that the Earth's field is of order 1 G!) These fields are strong enough to make the emission mechanism discussed above plausible. But as yet, there is no independent evidence for the strength of the magnetic fields; scientists can only estimate them from the spindown rate.

Pulsar magnetic fields fall into three groups. Normal pulsars have fields of order 10^{12} G. Faster-spinning **millisecond pulsars**, with periods below 10 ms, seem to have much weaker fields, of order 10^9 G. And there is a small group of very slowly rotating **magnetars** with fields of order 10^{15} G. These large differences in field presumably indicate different histories or modes of formation of stars in these classes. We will have more to say about millisecond pulsars below. Magnetars are in fact not normally seen in radio telescopes; they are found from the strongly pulsed X-ray emission.

The slowing down of the pulsar also allows us to estimate its age. If the slowing down were constant, so that the rate of change of the pulse period is constant for the whole lifetime of the pulsar, then we would have a simple equation

$$\text{period} = \text{original period} + \text{rate of change of period} \times \text{time}.$$

An estimate of the age of the pulsar is obtained by taking the original period to be zero (a neutron star spinning infinitely fast!) and solving for the time it takes to produce the present period with the observed rate of change of the period. If we let the rate of change of the period be represented by the symbol \dot{P} , which is the conventional symbol used by astronomers, then the age turns out to be $T = P/\dot{P}$. This is of course just an estimate. It could be an overestimate, since the original spin of the star was not infinitely fast. It could be an underestimate if the rate of change of the period was not constant in time, but was lower at earlier times. In fact, physicists expect that it was higher at earlier times, since the energy lost to radiation from the spinning magnetic field increases as the fourth power of the spin rate, so the spindown was much stronger when the pulsar was young. This leads to the conventional definition of the so-called "spindown age" of the pulsar,

$$T_{\text{spin}} = \frac{P}{2\dot{P}};$$

this is still, of course, an estimate.

For the Crab pulsar, the observed \dot{P} is 4.3×10^{-13} . (Notice that the rate of change of the period, \dot{P} , is a dimensionless number, because it is formed by dividing a number with dimensions of time – the change in the period – by another number with dimensions of time – the time in which the change took place.) From this, the spindown age is 1200 years. Astronomers in China happen to have recorded the supernova that produced the Crab in the year 1054, some 950 years ago. This is good agreement, considering the simplicity of the assumptions.

When astronomers use this method to compute the ages of other pulsars, they find that the Crab is the youngest known, and the oldest are some 10^9 years old. In fact, pulsar ages seem to correlate with their magnetic field strengths. The oldest pulsars, with ages up to 10^9 y, have millisecond periods and weak magnetic fields. Normal pulsars have ages up to about 10^7 y, although there are so many of them that it is not surprising to find a few with ages of a few thousand years. The magnetars are much younger, on average, than normal pulsars.

Investigation 20.2. Pulsars lose enormous amounts of energy

In terms of the total energy that a pulsar is getting rid of, pulsars are much more luminous than ordinary stars. The energy they lose as they slow down is enormous.

We can estimate that energy as follows. The energy of rotation is basically kinetic energy. But different parts of the star rotate with different speeds: the surface is going fastest and the center doesn't move at all. We shall approximate the rotational kinetic energy by assuming an em average speed: we won't be far wrong if we calculate the kinetic energy of a body with the mass of the neutron star moving with a speed that is half of the surface speed of the pulsar.

Consider the Crab pulsar. We can only do a rough calculation to see how large some of the numbers can be. Suppose the pulsar has a mass of $1M_{\odot}$ and a radius of 13 km, as our earlier calculations suggest is appropriate for neutron stars. Then the surface, rotating at 30 times per second, has a speed $v_{\text{surf}} = 2.5 \times 10^6 \text{ m s}^{-1}$, less than 1% of the speed of light. We take the average speed of the material in the star to be $v_{\text{avg}} = v_{\text{surf}}/2$, and so the kinetic energy of rotation K is

$$K = \frac{1}{2} M_{\odot} v_{\text{avg}}^2 = \frac{1}{8} M_{\odot} v_{\text{surf}}^2 = 1.5 \times 10^{42} \text{ J}. \quad (20.3)$$

Observations show that, in one period of rotation, the Crab pulsar's rotational speed decreases by a fraction 4.3×10^{-13} :

$$\frac{\Delta v_{\text{surf}}}{v_{\text{surf}}} = -4.3 \times 10^{-13}.$$

Exercise 20.2.1: Pulsar energy storehouse

A pulsar stores its energy as rotation. Estimate how much energy was released when the neutron star was formed by calculating the approximate gravitational potential energy of the neutron star, $-GM^2/2R$. You should find that the rotational energy is a small fraction of what was available when the star formed. What happened to the rest of the energy?

The energy K changes accordingly by the (negative) amount ΔK given by

$$K + \Delta K = \frac{1}{8} M_{\odot} (v_{\text{surf}} + \Delta v_{\text{surf}})^2.$$

Squaring the speed term on the right-hand side and subtracting the original expression for K given in Equation 20.3 above, we obtain

$$\Delta K = \frac{1}{8} M_{\odot} [2v_{\text{surf}} \Delta v_{\text{surf}} + (\Delta v_{\text{surf}})^2].$$

Now we divide by the original K to get

$$\frac{\Delta K}{K} = 2 \frac{\Delta v_{\text{surf}}}{v_{\text{surf}}} + \left(\frac{\Delta v_{\text{surf}}}{v_{\text{surf}}} \right)^2.$$

Since $\Delta v_{\text{surf}}/v_{\text{surf}}$ is so small, the second term on the right-hand side is completely negligible, and we have the simple result that the fractional decrease in the kinetic energy of rotation is twice that of the rotational speed itself, or -8.6×10^{-13} .

Now, the total kinetic energy is $1.5 \times 10^{42} \text{ J}$, so the loss of energy in one period of rotation is $1.3 \times 10^{30} \text{ J}$. Since one period takes only 0.033 s, this amounts to a rate of energy loss of $4 \times 10^{31} \text{ J s}^{-1}$. Compare this with the total luminosity of the Sun, about $4 \times 10^{27} \text{ J s}^{-1}$. The Crab pulsar is losing energy at the same rate as 10 000 Sun-like stars put together!

Puzzles about the rotation of pulsars

Some pulsars, particularly the youngest ones, show sudden small jumps in their spin rates, which astronomers call **glitches**. These seem to arise from the structure of the neutron star, whose interior is mainly fluid but which is thought to have a jelly-like **crust** of material formed from the heavy nuclei that neutron matter condenses into when the densities are below that of nuclei. Perhaps this crust breaks once in a while as the slowing of the star reduces its equatorial bulge, leading to a "starquakes". Or, perhaps the crust slows down a little more than the interior and once in a while has to be brought back up to the spin rate of the interior by the forces that connect the crust to the interior. Glitches are regularly seen in the Crab and other young pulsars.

By contrast, the older millisecond pulsars have not been observed to glitch. In fact they are excellent time-keepers. Some of them may even be better than the best atomic clocks made on Earth. This extraordinary stability has made them useful tools for conducting extremely sensitive observations that have verified the existence of gravitational waves and placed strict upper bounds on how much gravitational radiation has been left over from the Big Bang. We shall return to this subject in Chapter 22.

It may seem surprising that millisecond pulsars are old (as inferred from their very large spindown age) and yet they spin faster than the known young pulsars. This would indeed pose a big problem if there were so many millisecond pulsars that it seemed that all pulsars must turn into millisecond pulsars when they get old. But this is not the case. There are very few millisecond pulsars, especially considering that they remain visible for 10^9 y or more, while normal pulsars seem to stop radiating after about 10^7 y. Most old pulsars are probably slow rotators. The millisecond pulsar minority are formed in a special way, involving pulsars in binary

In this section: old fast pulsars, young slow ones, glitching pulsars, vibrating pulsars: pulsars puzzle astronomers more and more.

In this section: pulsars form in binaries but tend to split up the system and go their own ways. Sometimes they remain together, and some such systems are highly prized as laboratories where general relativity can be tested to high precision. One binary system has told us that gravitational waves are exactly as predicted by Einstein.

systems. We will return to that subject a little later in this chapter.

Pulsars in binary systems

Some of the most interesting pulsars are those in binary systems. Not many are known, which is consistent with our earlier discussion of binary disruption. But those that are observed are important in a number of respects. Significantly, almost all of them are millisecond pulsars, which suggests that pulsars that remain in binaries after being formed go on to become millisecond pulsars. Since they probably form in the same way as isolated pulsars (indeed, most isolated pulsars probably formed in binaries), something in the binary system must spin them up later. This probably involves the transfer of gas from the companion star onto the pulsar. This gas swirls around the neutron star as it gets near it, so when the gas reaches the star it carries considerable angular momentum, and it spins the star up. During the mass-transfer phase, the swirling gas will be hot enough to emit X-rays, and many systems like this have been discovered by X-ray telescopes in orbit. We will discuss this process in more detail in the section on X-ray binaries below.

Among the most significant pieces of information astronomers get from binary pulsars are their masses. Recall that all masses in astronomy are measured by studying orbits: the Sun's mass from the orbits of planets, the Earth's mass from the orbits of satellites, and so on. When pulsars orbit other stars, their orbits can tell us about their masses. Astronomers learn not just about the masses of the companions, but also about the masses of the pulsars themselves, since the pulsars are not simple test particles but instead produce observable effects on the motion of their companions.

Almost all the radio pulsars known to be in binary systems have companions that are either white dwarfs or other neutron stars. It may be that neutron stars in orbit about main-sequence stars have difficulty becoming pulsars, possibly because of the effects of gas coming from the companion. After the companion has evolved to a white dwarf, or has become a neutron star without disrupting the system, the binary is "cleaner", and the neutron star becomes a pulsar.

Pulsars give good information about the binary orbit because they are such stable clocks. The pulses are emitted by the pulsar at very regular intervals of time as measured by the pulsar itself, but their arrival times at the Earth change with the motion of the pulsar.

This is essentially the Doppler shift of the pulse rate, similar to any other Doppler shift. For binary systems, it is convenient to look at the Doppler effect in terms of the time interval between successive pulses, rather than in terms of the changes in the frequency of pulsation.

When the pulsar is in the part of its orbit nearest the Earth, the pulses take less time to reach the Earth than they do when the pulsar is on the other side of its orbit. The times between successive pulses therefore change in a periodic way as the pulsar orbits the companion. These changes are easily measured: the maximum delay between when a pulse might be expected (if there were no binary motion) and when it actually arrives is the light-travel time across the orbit, which can be several seconds. Measuring this effect tells us the size and period of the orbit, so it tells us something about the masses of the two stars.

The orbits of two-neutron star binaries are generally elliptical, again as expected from our binary disruption discussion. These are particularly fruitful, because, as we have seen in Chapter 18, elliptical orbits precess in general relativity: their orientation in space changes slowly. This can again be measured from the pulse train, since the exact pattern of delays depends on the orientation of the ellipse with

respect to the line-of-sight to the binary system. This gives a further constraint on the masses of the two stars and their separation.

Moreover, as the pulsar orbits the companion in an elliptical orbit, it finds itself sometimes nearer the companion than at other times. Just like any other clock, its pulsation rate experiences a gravitational redshift, which makes the pulses slow down and speed up periodically. Finally, the radiation from the pulsar follows a curved path as it passes the companion because of the deflection of light by the companion (see Chapter 18; discussed further in Chapter 23), and this introduces an anomalous time-delay into the arrival times of pulses that adds to the one produced by the redshift. Again because pulsars are such good clocks, the combination of these two small but changing effects can be measured, and it provides further information about the mass of the companion star and the size of the orbit.

The information in the combination of these measurements is enough to determine all the characteristics of the binary: the stars' masses, their separation, and even the angle that the orbital plane makes with the line-of-sight to the system. There are now two pulsar systems where the eccentricity is large enough and observations have continued long enough to make these measurements, and the results have given astronomers a surprise:

The masses of all four neutron stars in these two binaries are very close to $1.4M_{\odot}$.

There is mounting evidence from X-ray observations, which we will discuss below, that other neutron stars have masses about $1.4M_{\odot}$ as well.

It would seem that some mechanism operates to produce a fairly uniform mass, at least for neutron stars formed in binary systems. The mass of $1.4M_{\odot}$ is close to the Chandrasekhar mass for normal white dwarfs, as we saw in Investigation 12.5 on page 146, so it might be thought that the explanation is simple: only stellar cores above the Chandrasekhar mass will collapse, and they do so as soon as they reach that limit, so almost all neutron stars will have formed from cores of the same mass. The difficulty with this argument is that the mass is wrong: a white dwarf core that collapses with the Chandrasekhar mass of $1.4M_{\odot}$ will form a neutron star of about $1.2M_{\odot}$, because of the large amount of energy that is lost when it collapses. This energy loss is also a mass loss, by $E = mc^2$, and this can be as much as 10–20% of the original mass of the white dwarf. So a $1.4M_{\odot}$ neutron star must have come from the collapse of a core with a mass of about $1.6M_{\odot}$, or some extra mass must have been added to the star after the initial collapse. It is not yet clear what this means about the circumstances in which these stars form.

X-ray binary neutron stars

Not long after the discovery of pulsars, astronomers found other neutron stars in an equally unexpected place: X-ray binary systems. We saw above that one might expect to detect X-rays from neutron stars, because at least young ones can be very hot. But the first X-ray satellite observatories did not see the expected relatively weak black-body X-radiation from points located in the middle of supernova remnants. Instead, they found powerful X-ray sources scattered over the sky, and it soon became clear that at many of these locations there was also a visible giant star whose spectrum showed periodic Doppler shifts. These sources were evidently binary stars which somehow produced huge amounts of X-rays.

Assuming that the X-rays were black-body radiation, the sources had to have high temperatures. Moreover, luminosity of the sources required much larger surface areas than the area of a neutron star. The temperatures could not be supplied by the visible star, whose surface temperature could be measured from the observed

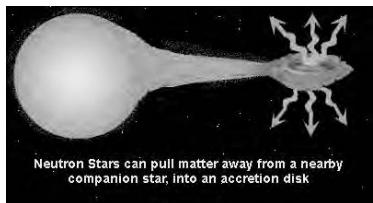
>Once the masses of the stars in a binary system have been determined, it is possible to make one further prediction: how much energy should be radiated by the orbital motion in gravitational waves. We will see in Chapter 22 how to calculate the radiation emitted. But in general terms, it is clear that the effect of losing energy must be to gradually shrink the stars' orbits, bringing them closer together. As the orbit shrinks, the orbital period goes down as well. This should be noticeable in radio observations, and indeed in two systems this effect has now been seen. We will discuss the importance of this for understanding gravitational radiation in Einstein's theory in Chapter 22.

In this section: some binary systems with neutron stars are spectacular emitters of X-rays. These systems convert mass into energy far more efficiently than any nuclear power station or explosive made by man.

visible light. They had to be associated with its binary companion, which was not visible in optical light. It was significant that the phenomenon usually seemed to involve giant stars, stars that were at a point in their evolution where they had suddenly expanded greatly in size.

These facts soon led astronomers to the now-accepted **accretion** model of binary X-ray sources. The unseen companion that is the source of the X-rays is a neutron star or black hole, and it is close enough to the visible star that when the visible star expanded to become a giant, it began to dump mass onto the compact star. This happens in some cases because the giant star begins blowing off mass at a considerable rate (a giant version of the solar wind), and in other cases because the giant expands so far that the part of it nearest to the neutron star actually begins to be dominated by the gravity of the neutron star, so that a steady stream of gas falls towards the neutron star.

Figure 20.5. An artist's sketch of accretion onto a neutron star in a binary system. Because the neutron star is in orbit about its companion, the matter that falls onto it is rotating, and forms a flattened disk. Drawing courtesy NASA.



In either case, because of the orbital motion of the compact star, the gas does not fall onto it in a spherical manner. Instead, it first swirls into a disk in the plane of the orbit, and then the gas gradually spirals down onto the compact star. As it spirals towards the star, it releases its gravitational energy. The spiralling-in must happen due to friction within the gas, so the released energy goes first into thermal motions in the gas, and then into black-body radiation. The radiation that we see is coming from the disk, not from the compact star: the disk has much larger surface area and therefore radiates much more energy. The disk is called an accretion disk, since it contains the gas that will accrete (accumulate) onto the central star.

The compact star must be at least as compact as a neutron star: if it were a white dwarf, the temperature of the disk would never get high enough to make X-rays. We show this when we study accretion disks in Chapter 21. Because the gravity of a neutron star is so strong, the amount of energy released as a particle moves through the disk is a large fraction of its rest-mass, up to 10–20%.

Using strong gravity as a mediator, accretion disks convert a much larger fraction of their rest-mass into energy than do the nuclear reactions inside stars!

With this much energy available, it is possible to power X-ray sources with very little mass: the amount of mass accreted by a neutron star can be as little as $10^{-9} M_{\odot}$ per year. Since the giant star may remain a giant for only a few million years, the total amount of mass dumped onto the neutron star need be only a small fraction of a solar mass. But this much mass is enough to change the spin of the neutron star substantially during this time: it is easily possible to spin stars up to 600 Hz as they accrete the rotating gas from the disk.

We shall return to accretion when we consider black holes. There we will put more numbers into the discussion above.

Gamma-ray bursts: deaths of neutron stars?

In the 1960s, the United States Air Force launched the first of a series of Vela satellites designed to detect tests of nuclear weapons in space that had been banned by treaty. Later Vela satellites contained gamma-ray detectors and good clocks, so that the position of an explosion could be determined by differences in the arrival time of the gamma-rays at different satellites. These satellites never detected violations of the test-ban treaty, but in the early 1970s Air Force scientists realized that over

In this section: gamma-ray bursts, the most energetic events astronomers observe, may be associated with neutron stars that are on their way to becoming black holes.

the years a number of events had been detected that had come from astronomical sources.

Since astronomers could not find any other wavelengths of light associated with these **gamma-ray bursts**, they could not determine their source. It was not even known *where* they were coming from. But as data accumulated, especially from the 9-year mission of NASA's Compton satellite (launched in 1991), it became clear that the bursts were at least as far away as remote portions of the Galaxy, if not further, and that they must be coming from objects as compact as neutron stars. The principal reason for this was the time-scale of the bursts. Each burst was highly individual, but most lasted from about 1 s to several hundreds of seconds. Moreover, each burst was composed of sub-bursts where the intensity changed dramatically on times as short as a millisecond. As we noted above, this pins them down to neutron stars.

Finally in the mid-1990s astronomers, using other satellites, managed to find visible light emitted by the bursts. For a few hours to days, the region around the burst would brighten with what astronomers call an "afterglow". Bursts were always found to be located in distant galaxies. In fact, there seemed to be no limit on how far away they could be seen. They were bright enough to be seen as far away as any galaxy. During the burst the gamma-rays were thousands of times as luminous as the entire host galaxy. Since then, bursts have been seen that rival the luminosity of the entire Universe, for the brief seconds that they shine. Such an enormous energy can only be obtained by converting a good fraction of a solar mass into pure radiation. Given that it happens on short time-scales, this means that something catastrophic is happening to a neutron star or black hole. Bursts happen several times a day somewhere in the Universe.

Astrophysicists have come up with various proposals for what happens in a gamma-ray burst. As we have seen, neutron stars sometimes form binary systems, whose orbits shrink gradually because of the emission of gravitational radiation. What happens when an orbit shrinks so much that the stars are brought together? They cannot form a new neutron star, since the maximum mass is thought to be less than the $2.8M_{\odot}$ that the two stars contain. (See the discussion of this below.) They might quietly collapse to a black hole, but this does not seem likely, given the speed with which they collide and the highly distorted shape that they have when they first encounter one another. It seems more likely that they explode, at least partially. This explosion should produce visible X-radiation, and it might even produce a gamma-ray burst.

Another possibility is that the burst occurs at the end of the in-spiral of a binary system containing a neutron star and a black hole. While no such systems have yet been identified in radio pulsar observations, they should exist. In this case, the neutron star is disrupted by the tidal forces of the black hole and must form a thick ring close to the hole, which loses energy and spirals into the hole. In this process, it may well be possible to expel some of the energy in a jet of gas that gives rise to the burst.

Some astrophysicists have proposed that the event takes place inside a very massive star, as a kind of super-supernova, called a **hypernova**. It might still involve the formation of a neutron star at an intermediate stage, but this star does not live long before forming a black hole and releasing further energy.

Whichever of these models is correct, if any, the energy and timing of the burst strongly suggest that this is the catastrophic death of a neutron star. The energy comes, in the final analysis, from the gravitational energy released when matter forms objects as compact as neutron stars or black holes, as we calculated in Exer-

►The Compton satellite was named after Arthur Compton, discoverer of the scattering of photons by electrons, whom we met in Chapter 8. It used Compton scattering of gamma-rays to detect them.

Investigation 20.3. Building neutron stars in general relativity

Despite the complexity of general relativity, it is possible to cast the equations for the structure of a star in a form that is very similar to the Newtonian equations. It then becomes simple to modify our program Star to produce such models with little effort.

Let us remind ourselves of the Newtonian equations. We imagine stepping outwards in radius with steps of size h . The mass of the star increases according to Equation 8.16 on page 94:

$$m(r+h) = m(r) + 4\pi r^2 \rho(r) h.$$

The pressure decreases by the equation of hydrostatic equilibrium, a combination of Equation 8.17 on page 94 and Equation 7.1 on page 73:

$$\Delta p = -G \frac{\rho m(r)}{r^2} h.$$

In relativity, it turns out that the mass equation above remains exactly the same. The relativistic corrections to the hydrostatic equilibrium equation all involve terms proportional to $1/c^2$, which would be very small in the non-relativistic limit, where c is large compared to any physical speed. The new equilibrium equation must be calculated using the full mathematical framework of general relativity, but when that is done the result is remarkably simple. There are only three changes, and each of them can plausibly be related to a property of relativity that we have already encountered. They are as follows.

1. Replace

$$\rho \rightarrow \rho + p/c^2.$$

This is the inertial mass density, defined in Equation 15.10 on page 193, which governs how a fluid accelerates under an applied force.

2. Replace

$$m(r) \rightarrow m(r) + 4\pi r^3 p/c^2.$$

This is analogous to the active gravitational mass. In Newton's theory, the density of the active gravitational mass is just ρ , and $m(r)$ is the total mass inside a radius r . In relativity, the active mass density is $\rho + 3p/c^2$ (for a fluid with isotropic pressure, as we have assumed here). If we multiply the extra term, $3p/c^2$, by the volume inside radius r , $4\pi r^3/3$, then we just get the extra term in the relativistic structure equation. Now, at one level this "derivation" is just a coincidence, because the actual volume of the star inside the radius r is not given by the flat-space formula, since space is curved. And the total pressure "inside" radius r would not be obtained by multiplying the value of p at r by this volume, because p is larger at smaller radii. But the coincidence is nevertheless interesting, and it does make it clear that this term does give the role played by the active gravitational mass in the equation governing the structure of the star.

3. Replace

$$r^2 \rightarrow r^2(1 - 2Gm(r)/c^2r).$$

This is related to the curvature of the three-dimensional space of the star. Recall from Chapter 19 that, at least for weak gravitational fields, the coefficient of $(\Delta x)^2$ in the spacetime-interval is $1 + 2\Psi$, where, to our accuracy, Ψ is the same as the Newtonian field $\Phi = -Gm(r)/c^2r$. So the square of the change in the proper distance $(\Delta l)^2$ between two points is $(\Delta x)^2(1 - 2Gm(r)/c^2r)$ near a point at radius r . This is close to the form of the new term in the equation, but not exact. Again, the analogy is enough to help us to see that this term in the equation is an effect of spatial curvature.

The result of these replacements is the relativistic equation of hydrostatic equilibrium, called the *Oppenheimer-Volkov equation*,

$$\Delta p = -G \frac{(\rho + p/c^2)(m(r) + 4\pi r^3 p/c^2)}{r(r - 2Gm(r)/c^2)} h. \quad (20.4)$$

We have used this equation in the program Neutron.

A few cautionary words are in order about the interpretation of some of the symbols. We have said in the text of this chapter that physical variables such as p and ρ are measured by a locally freely-falling experimenter at rest in the fluid. But r and $m(r)$ are not defined this way, since they are not locally measurable. The radius r has a well-defined geometrical meaning, even in the curved spacetime of a star. Since the spacetime is spherically symmetric, there are surfaces around the center of the star that are perfectly spherical. The coordinate r assigned to any surface is the circumference of the sphere divided by 2π , just as in flat space. This may not seem worth commenting on, until one realizes that the definition makes no mention of the distance to the center. In a spherical but curved space, this distance will generally not be equal to r . We have already noted this in our discussion of the way the denominator of the equilibrium equation changes. We shall see an example of this in Chapter 21.

The mass $m(r)$ is best regarded as an auxiliary variable. It cannot be interpreted as the mass inside the sphere whose radius is r , as it would be in Newtonian gravity, since (1) mass is not easy to localize in relativity, especially because gravitational potential energy has to be counted in the total mass but does not reside in any particular place, and (2) the volume inside radius r will not in general be the same as in flat space, so the factor of $4\pi r^2 h$ in the mass equation is not the actual volume of the shell of thickness h . In practice, this does not cause a problem: one never needs to know how much mass is within a certain radius of the star.

Finally we need to discuss the equation of state. We shall only use the non-relativistic neutron-gas equation of state, despite the fact that the neutrons do become relativistic when the density goes much beyond the normal density of nuclear matter. The reason for neglecting relativistic corrections is that, at the point where they become important, other aspects of nuclear physics are so poorly understood that the relativistic corrections will by themselves not be very accurate. It is important to understand, however, that the results of our models at high densities are only indicative, and cannot be relied on quantitatively.

The non-relativistic equation of state is of the form given by Equation 12.14 on page 144:

$$\rho_{\text{Fermi}} = \beta \rho^{5/3}, \quad (20.5)$$

where the argument in Investigation 12.4 on page 144 does not determine the constant β accurately. We merely quote the correct coefficient, which can be derived by careful calculations like those outlined in Investigation 12.4:

$$\beta = \left(\frac{9h^6}{320\pi^2 m_n^5} \right)^{1/3},$$

where in this equation h is Planck's constant. In SI units this evaluates to

$$\beta = 5.3802 \times 10^3 \text{ kg}^{-2/3} \text{ m}^4 \text{ s}^{-2}.$$

This value is used in the program Neutron on the website.

The program to evaluate the models is essentially the same as that used for the Sun, the program Star, also on the website. Only one difference needs to be commented on here: the irrelevance of temperature. The equation of state we use has a fixed coefficient β , given by the physics of a degenerate neutron gas. It is not affected by the temperature, since in a degenerate gas the pressure comes from quantum effects and not the random motions of the atoms. So not only is temperature not important to the structure of the star, it is also wrong to calculate it from the degenerate pressure using the ideal gas law. So we leave it out of the program entirely.

Other details of the program are explained on the website.

cise 20.2.1 on page 275. This energy is so large, and it is released so quickly, that it makes the nuclear reactions going on in ordinary stars seem insignificant. Fortunately for the Earth, our Sun relies only on the tame nuclear physics!

The relativistic structure of a neutron star

Now we return to the theory of neutron stars. Neutron stars of a solar mass are highly relativistic, as we have seen. It is not reasonable to expect that Newtonian stellar models, of the type we made for ordinary stars and white dwarfs, would be very accurate here. In fact, we shall see that they are more than inaccurate: Newtonian models completely miss some very important features of neutron stars.

The structure of a neutron star in general relativity is affected by a number of things. First, of course, there is the curvature of spacetime, and particularly that of space itself. Then there are the effects of special relativity in its equation of state. And finally, one has always to be careful about how one defines the quantities one deals with, since different observers may see them differently.

A good example of the problem of definition is given by the mass density ρ . First of all, it must now include all energies, since all energies have an equivalent mass (the energy divided by c^2) and they all contribute to the inertia of an element of the fluid. Second, this mass(-energy) density will depend on the observer. Different observers measure different energies for any body, and the fluid in the star will be no exception. On top of that there is the Lorentz–Fitzgerald contraction (see Chapter 15), which means that an observer moving past a little piece of the fluid will measure its volume to be smaller, and hence its density to be larger. So what do we mean by the *density* of the gas in a neutron star?

In order to avoid confusion over a multitude of possible definitions, physicists have settled on the following convention. When they use a word in relativity that is the same as one in Newtonian physics that refers to a property of a gas, such as density or temperature, they mean the analogous quantity *as measured by a freely-falling experimenter who is at least momentarily at rest in the part of the fluid which is being measured*. Being freely-falling, the observer can make a measurement as if there were no gravity, as if the whole system were governed just by special relativity. Since the experimenter is at rest with respect to the fluid, the effects of relativistic speeds do not come into it.

It must be stressed that this is a convention on the definition of symbols like ρ and T . There is of course real physics in the interactions between fluid streams that move with respect to one another, and between the gravitational fields they produce. An observer moving through a fluid might try to measure the density, and would not be likely to get ρ , since that is the density measured by an observer at rest. It is the job of general relativity to deal with these things, and it does it very well. But when we want to describe the state of a fluid, we will always use the measurements that an experimenter at rest inside it would make.

We cannot derive here all the differences between the equations that govern the structure of a star in relativity and those in Newtonian gravity. Despite the complexity of Einstein's equations, it is remarkable that the equation of hydrostatic

In this section: we use general relativity to make realistic computer models of neutron stars. The equations governing their structure differ from those for Newtonian stars in ways that correspond directly to the new features of relativistic gravity that we learned about in previous chapters.

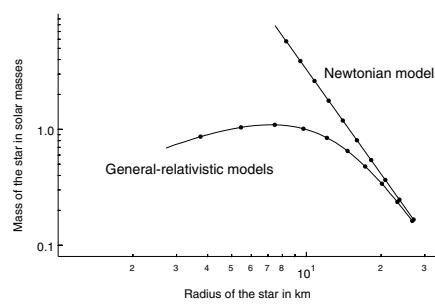


Figure 20.6. The mass–radius relation for neutron stars constructed in both general relativity (using Neutron) and Newtonian gravity (using Star). In both cases the equation of state is the non-relativistic form of the neutron-gas equation of state that was derived in Chapter 12. The data show that general relativity predicts a maximum mass for neutron stars even when Newtonian stars using the same internal physics have no maximum mass.

equilibrium can be written in a form that is so similar to the one we have in Newtonian mechanics. We describe this equation in Investigation 20.3 on page 280, where we also go on to construct models of neutron stars using the computer program **Neutron** on the website, which is a modification (and in some ways a simplification) of the program **Star**.

The results of several runs of **Neutron** and **Star** are shown in Figure 20.6 on the previous page. This figure shows the main features of how relativity affects neutron stars, so we shall discuss it here at some length.

The relation of mass to radius for neutron stars

In this section: general relativity sets a maximum on the mass of neutron stars. This comes from the fact that pressure creates gravity. Fortunately for life on Earth, this maximum mass is *larger* than the Chandrasekhar mass.

Figure 20.6 on the preceding page compares the mass and radii of models constructed within relativity and Newtonian gravity, starting with the same central density and using the same equation of state. The differences between the curves are striking, and they are due entirely to the relativistic corrections to the Newtonian structure equations. The differences range from 2% for low-density stars up to more than 70% for the very dense models.

The most obvious difference between the two sequences is that the relativistic stars have a *maximum mass*. The value of this maximum is too low, a little lower than the measured mass of binary neutron stars. This simply means that we are not using the right equation of state, and we know that already: we have not put in relativistic corrections for the neutron gas, and on top of that we are not modeling the hard-core repulsive part of the nucleon–nucleon force at all. This hard core should increase the pressure at high densities and raise the maximum mass considerably. So the figure shown here is instructive more for what it tells us about general relativity than about real neutron stars.

And it tells us that general relativity will place a maximum on the mass of a sequence of stars, even when there is no such maximum in Newtonian sequences. In fact, the Newtonian sequence moves along a perfectly straight line in this logarithmic plot, which means that the relation between mass and radius for Newtonian stars is a power-law: $M \propto R^\alpha$. This is an illustration of a fact that we used but did not prove in Chapter 8, that in Newtonian gravity a polytropic model has a perfect scaling: if one knows its structure for a certain central density and temperature, then one can deduce its structure for other central values by simple scaling.

The relativistic equations do not obey this scaling: every model is different from every other one, even for a polytropic equation of state.

The maximum of the mass has a fundamentally important consequence. Any collapsing cloud of gas with a mass larger than this maximum cannot stop at neutron star density; it must continue collapsing, presumably to a black hole. This makes the formation of black holes all but inevitable in astronomy.

The only way to avoid black holes is for nature to insure, magically, that collapsing clouds do not exceed this mass. But this does not seem plausible. The maximum mass is a property of the very dense neutron-matter equation of state, and should not be related to processes that lead to collapse in giant stars. Moreover, giants extend in mass up to many tens of solar masses. It seems unlikely that in all cases the star can insure that no more than, say, two solar masses of material actually collapses.

The existence of a maximum has another important but less obvious consequence: it tells us that all the neutron star models with a mass smaller than the maximum but with a central density that is higher than that of the maximum-mass star are *unstable* and will themselves collapse if disturbed.

The reasoning is similar to that used in Investigation 8.8 on page 101. Imagine starting with the star which has exactly the maximum mass, and then constructing another model with a slightly larger central density. Since the mass is a maximum, the new model has essentially the same mass as the previous one, but it has a larger density and therefore a smaller radius. Being an equilibrium model, it represents a fluid configuration that will remain static indefinitely. Physically, the structure of the star is the same as if we had taken a single star and compressed it: its mass would not change but its central density would go up and its radius would decrease. Since our calculations tell us that after these changes we still have an equilibrium model, we learn that the model with maximum mass is *neutrally stable*: if it is compressed it will neither bounce back out or collapse – it will just remain in equilibrium at the smaller radius.

Now, this neutral stability is a very special property: it indicates a point of transition between stability and instability along a sequence of equilibrium models. Consider a hypothetical sequence of stars which makes such a transition. On one side of the transition point, the stars are stable: they respond to a compression by bouncing back. On the other side they are unstable; they respond by collapsing further. The transition point must therefore be a point where the response is to remain exactly in equilibrium. Therefore, when we build a sequence of stars, we need only make a mass–radius plot and look for places of maximum (or minimum) mass: these are places where stability changes along the sequence.

So we only need ask which side of the point of maximum mass is the stable side and which is the unstable one. The models to the right have lower density, and the very low density ones are very similar to the Newtonian models. The polytropic exponent γ that we have used here is $5/3$, which we saw in Chapter 8 means that the Newtonian models are all stable. It follows that the relativistic ones far to the right of the transition point are also stable. Moving from these to higher central densities, the stability cannot change until the mass reaches a maximum or minimum, so we conclude the following.

All the relativistic models in Figure 20.6 on page 281 to the right of the maximum neutron star mass are stable, and all the high-density models to the left are unstable.

If we had used a better equation of state at high densities we would have seen the same general feature, with the mass reaching a maximum, but the mass would be higher and the peak at maximum might be sharper. I have not bothered to introduce a more realistic equation of state because at present there is still considerable uncertainty about the details at high densities.

Neutron stars as physics labs

The uncertainty we presently have in the equation of state is one motivation for studying neutron stars: observations have the potential to tell us things about nuclear physics. Good measurements of the radius as well as the mass of a neutron star, or some indication of what the actual maximum mass is, would bring a rich reward in nuclear physics.

The physics of the interior of a neutron star has many more challenges than just computing the correct equation of state. Here are a few.

Neutron stars clearly have strong magnetic fields, which lead to pulsar emission, but physicists do not know what creates and maintains the magnetic field. It may be similar to the processes that maintain the Earth's field and that of the Sun (which are also not very well understood), or it might be totally different. Any explanation

▷ This argument only tells us about stability against *spherical* disturbances. We shall not consider non-spherical ones here.

In this section: neutron stars contain matter in extreme conditions. Their interiors are as yet veiled from observation, but they probably contain superfluid and superconducting material, even at temperatures of millions of degrees. They may have jelly-like crusts and solid cores. Future observations of neutron stars may reveal even more exotic physics.

must also explain why the magnetic poles lie near the equator of a neutron star, whereas for the Earth and the Sun the poles are near the rotation poles.

The neutrons in the deep interior are thought to form a what physicists call a superfluid. This is a fluid that moves with no **viscosity**, no friction. Whether or not this happens depends on details of the nuclear physics that are not testable in laboratory experiments. Superfluidity is a peculiarly quantum phenomenon; it arises from the fact that the neutrons are all identical to one another. Under some conditions, they all move in exactly the same way, and it takes a lot of energy to make them scatter from each other or from other particles, because to do so would place the scattered particles into a quantum state where they can be distinguished from the others. The indistinguishability of the particles also means that they cannot rotate about one another, because again this would mean that the particles at the center of rotation would be distinguishable from the others. Superfluids can only rotate about special lines called **vortices**, within which the fluid is not a superfluid. But neutron stars do rotate, so they must be threaded with enormous numbers of these vortices. Physicists understand little about these vortices, or the way they interact with the magnetic field whose axis runs perpendicular to them.

The protons in the interior may also form an analogous fluid, a superconductor. This has no electrical resistance. This might or might not be related to the mechanisms that create the magnetic field.

At the very center of the star the density may be high enough to allow some even more exotic physical processes. These may involve elementary particles that are hard to study in the lab, such as **quarks**, which are the building blocks from which protons, neutrons, and other strongly interacting particles are made. The central core is thought to be liquid, but under some equations of state it may be solid.

Our understanding of nuclear physics is not good enough today to exclude the possibility that there is another stable state of matter even denser than nuclear matter. This would be called **quark matter** or **strange matter**, and it is possible that this is the most stable state of matter, leading to **strange stars**.

The outer parts of the star, where the density is low enough that not all matter has turned into neutrons, consist of layers of very unusual neutron-rich isotopes of familiar elements. The properties of the semi-solid outer crust – how it wobbles, breaks, re-forms, etc. – depend on these nuclei and how they fit together into regular patterns. The vortices of the interior rotation terminate on the inside of the crust, which is not superfluid, so it rotates normally. Slippage of these vortices along the crust is one model for the glitches in young pulsars.

Scientists today are only able to give relatively superficial descriptions of this physics; they do not have enough experimental data to help them to understand the details of these fascinating objects. It is not possible simply to see inside the stars. Data from supernovae that form neutron stars will help. More helpful will be the observation of gravitational waves from neutron star vibrations (Chapter 22), because then neutron star seismology could begin to reveal the inner structure in the same way that helioseismology has shown us the inside of the Sun (Chapter 8). However, it will probably be many decades before astronomers can gather enough data to unravel the mysteries of the structure of a neutron star. Until then, neutron stars will remain the most mysterious stars in the Universe. Only the interiors of black holes are more effectively hidden from our view.

Black holes: gravity's one-way street

Black holes. No term evokes the mystery of modern gravity more than this one. The mystery of black holes is more than an invention of popularizers of astronomy and relativity. Black holes were certainly a mystery to Einstein and his contemporaries. Yet today black holes are everywhere: in X-ray binaries, in the centers of galaxies, and of course in books, like this one, on relativity and gravity!

Theorists attacked the problem of understanding black holes, not by using astronomical evidence, but by using lessons they had learned from quantum mechanics. Quantum thinking demanded that physicists ask only questions about things that could be measured, not about what is hidden from experiment. Thus, they can measure that light behaves sometimes as a particle (the photon) and sometimes as a wave, but they find it useless to ask what *is* a wave-particle.

In quantum thinking an object *is* only what it *does*.

Theorists found this disciplined way of thinking helpful for relativity too. It meant that one should only ask what one can measure about black hole, and not worry about the rest. The rest includes, for example, strange things that might happen to coordinates. If a distant observer or an experimenter falling into a black hole can measure something, then it is real and important. If not, then forget it.

This rule is a good one for learners of relativity too. Always frame questions in terms of what an observer could measure. If it is impossible to frame the question that way, then it is not a question that has any real meaning.

Many physicists contributed to developing this point of view, but none has been a stronger advocate of it than the American physicist and teacher John Archibald Wheeler (b. 1911, see Figure 21.2 on page 288), who, incidentally, was the person who coined the name "black hole".

The first black hole

Remarkably, the black hole was the very first exact solution of Einstein's equations that physicists found! Einstein had contented himself at first with approximate solutions that made key predictions, like the bending of light by the Sun and the precession of the perihelion of Mercury (Chapter 18). But the German astronomer and theoretical physicist Karl Schwarzschild (1873–1916) almost immediately found the solution describing exactly the gravitational field outside a spherical star.

Here is the spacetime-interval Schwarzschild found:

$$\Delta s^2 = - \left(1 - \frac{2GM}{c^2 r}\right) (c\Delta t)^2 + \left(1 - \frac{2GM}{c^2 r}\right)^{-1} (\Delta r)^2 + r^2 [(\Delta\theta)^2 + \sin^2\theta(\Delta\phi)^2]. \quad (21.1)$$

In this chapter: we study general relativity's most intriguing prediction: black holes. We look at the central place they have in Einstein's theory, their role in astronomy today, and the direction they are giving to efforts to unify gravity and quantum theory. We calculate orbits around black hole, examine the astronomical evidence for black holes, and learn about wormholes, the Hawking radiation, and black hole entropy.

>The image under the text on this page is a spacetime drawing of the light-cones near a black hole. It is a thin slice of time, showing in which directions light can travel after being emitted from different locations in space near the black hole. Far from the center, the cones are vertical, and they open 90°, as in flat Minkowski spacetime. But near the center, they tilt more and more inwards and get narrower. Image courtesy W. Benger, ZIB and AEI.

In this section: the Schwarzschild black hole was the first solution found for Einstein's full equations, but one of the last to be understood. Besides describing a black hole, it gives the gravitational field outside any spherical (non-rotating) star.



Figure 21.1. Karl Schwarzschild was one of the most brilliant astrophysicists of his time. A friend of Einstein and the director of the astronomical observatory in Potsdam, near Berlin, he was serving with the German artillery on the Russian front in the First World War when he received copies of Einstein's papers announcing general relativity. Two months later, in January 1916, Schwarzschild sent Einstein the solution, which Einstein immediately had published. Tragically, four months later Schwarzschild was dead, killed by disease. One can only speculate on what he might have accomplished had he survived the war. In particular, he would likely have grasped the importance of Chandrasekhar's maximum mass for white dwarfs, and this could have led to a much earlier recognition of the astrophysical importance of neutron stars and black holes. Schwarzschild's son Martin (1912–1997) later fled Hitler, settled in the United States, and became one of the pioneer developers of the theory of stellar evolution. Image reprinted with permission of the Universitäts-Sternwarte Göttingen.

In this section: the Schwarzschild solution exhibits strong gravitational redshifts. A photon can stand still at the horizon of the black hole.

In this equation there is a constant named M , which is the mass of the black hole. Notice that this constant enters in combination with G and c , so that the equation actually contains only the length GM/c^2 . We first met this length in Equation 4.12 on page 36, where we called it the *gravitational radius* R_g associated with the mass M .

Although the whole expression may at first look complicated – it is probably the most complicated equation we will write down in this book – it can be understood if one approaches it by asking questions about measurables. We take it apart in Investigation 21.1 and see what its pieces say. But one piece is so important that we should treat it here: the piece governing the curvature of time.

We know that the time part of the spacetime-interval gives the rate at which clocks at a fixed place in space run. If a clock is at rest, so that along its world line $\Delta r = \Delta\theta = \Delta\phi = 0$, then the Schwarzschild spacetime-interval tells us that its proper time lapse $\Delta\tau$ associated with a coordinate time lapse of Δt is given by

$$(\Delta\tau)^2 = -\frac{1}{c^2}\Delta s^2 = \left(1 - \frac{2GM}{c^2r}\right)(\Delta t)^2. \quad (21.2)$$

This determines the curvature of time, as we saw in Chapter 18.

Remarkably, the time part of the spacetime-interval is exactly the same as in Equation 18.7 on page 231. This leads to three conclusions. First, if we are far away from the center of the Schwarzschild geometry (i.e. if r is much larger than the gravitational radius GM/c^2), the curvature of time is the same as it is far from a nearly Newtonian star whose mass is M . We saw in Chapter 18 that the curvature of time determines the orbits of planets and other slowly moving objects, so we are led to the following important conclusion.

Gravity far from a black hole is just like gravity outside an ordinary star: you can't tell you are in orbit around a black hole just by measuring the orbits far away from it.

Second, we are right to call the constant M in the Schwarzschild geometry the “mass” of the black hole. It is the mass that an ordinary star would have if it produced the same Newtonian gravitational effects far away. Third, we have learned that the coordinate time t is the time as measured by experimenters far away, just as in Chapter 18.

At first, no-one knew that the Schwarzschild solution described a black hole. As a description of the geometry outside an ordinary star, it is just a generalization of Equation 18.15 on page 235. But suppose that the star is smaller than $2R_g$, so that the point $r = 2R_g$ is outside the star. Then we can presumably use the interval in Equation 21.1 on the preceding page at this value of r , but of course something strange happens there. The coefficient of $(\Delta r)^2$ in the spacetime-interval in this equation gets infinitely large as r approaches $2R_g$, so that it seems that proper distances stretch out infinitely far. And as Equation 21.2 shows, time seems to stand still there. We call this special radius $2R_g$ the **Schwarzschild radius**. To find out what happens at this radius, we have to ask about measurable things: what do black holes *do*?

What black holes can do -- to photons

The bad behavior at the Schwarzschild radius was the problem that took such a long time for physicists to solve. As we noted before, the way to understand these things is to ask questions strictly about measurable things. For example, how shall we understand that time seems to stand still at $r = 2R_g$?

Investigation 21.1. The Schwarzschild geometry

The Schwarzschild geometry is so important that it is useful to have a look at it. It is not hard to understand it, after the spacetime-intervals we have already studied.

Here, copied from Equation 21.1 on page 285, is the Schwarzschild spacetime-interval:

$$\Delta s^2 = - \left(1 - \frac{2GM}{c^2 r}\right) (\Delta t)^2 + \left(1 - \frac{2GM}{c^2 r}\right)^{-1} (\Delta r)^2 + r^2 [(\Delta\theta)^2 + \sin^2\theta (\Delta\phi)^2]$$

The first things to notice are what is *not* in the spacetime-interval, i.e. how simple it is. One “not” is time: the coefficients are all independent of time. Time appears in the spacetime-interval as Δt , of course, but the *coefficients* of the coordinate changes are all independent of t . This tells us that the geometry described here is time-independent. It represents the gravitational field outside a star that is simply sitting there, doing nothing!

Another thing that is missing is any mixed term, such as a term with $\Delta t \Delta\phi$. From our discussion of the dragging of inertial frames in Chapter 19, we would expect such terms if the star were rotating. Therefore, this geometry is that outside a static star, one that has no internal fluid motions.

Now let us look at what *is* in the spacetime-interval. The last term in square brackets is just the distance relation on a sphere of radius r , as we worked out in Equation 18.4 on page 228, so this expresses the fact that the geometry is spherical. More precisely, the two-dimensional surfaces on which t and r are constant (obtained by setting Δt and Δr to zero in the spacetime-interval above) have the same geometry as a sphere of radius r in flat space.

Now, in flat space r would be the distance to the center of the sphere. In this spacetime-interval, however, the term in square brackets has nothing to do with the distance to the center of the sphere, which is measured by the Δr part of the spacetime-interval. What it tells us is that, if we stay on the sphere, then little steps in angle require the same distance as little steps on a sphere in flat space.

with a radius of r . In particular, if we add up a lot of little steps and go all the way around the sphere, we will measure a circumference of $2\pi r$, just as in flat space.

So the coordinate r is a “circumferential radius”, defined by how broad the sphere is, not how far it is to the center. In a curved space the circumference does not have to equal 2π times the radial proper distance. So when writing down the spacetime-interval, we must always make choices about what coordinates to use; this is a point we made earlier. Here we have a coordinate system whose radial coordinate is defined by circumferences of spheres. Later we will meet a different coordinate system for this geometry.

The spherical part tells us how the radial coordinate is defined and how it can be measured (by measuring the circumference of a sphere), but it is not a surprise that it is there: we know Schwarzschild assumed a spherical geometry from the start. More interesting are the coefficients of the time part of the spacetime-interval and the radial part. We discuss the time part in the main body of this chapter. So let us turn to the spatial part.

The spatial curvature is determined by the coefficient of $(\Delta r)^2$. Looking at the expression, it appears that it is everywhere larger than one. This measures proper distance in the radial direction, so one can say that radial distances are bigger than we would find in flat space: two spheres with circumferences $2\pi r_1$ and $2\pi r_2$ are separated by a proper distance that is *larger* than $r_1 - r_2$. This is the effect of the curvature.

Why not go all the way to the center of the sphere: what is the radial distance? There is a problem with this question, because the coefficient of $(\Delta r)^2$ goes to infinity at the finite radius $2R_g = 2GM/c^2$. We explore the meaning of this difficulty in the main body of this chapter.

One of the points of genius of Schwarzschild’s solution is his choice of the radial coordinate. He saw clearly how useful it would be to have a definition of the radius of the spheres that was not tied to a notion of radial distance. His r can be defined without reference to the center of the space, which could be buried in the middle of a star.

Recall that the coordinate time t is the time on a distant clock. We saw as early as Chapter 2 that time on a clock slows down relative to distant clocks as the clock goes deeper into a gravitational field. This is the effect of the gravitational redshift. So the fact that proper time on a clock should be smaller than coordinate time t , as in Equation 21.2, is to be expected. It just means that, if the clock sends out one photon every second by its own proper time, then the deeper into the gravitational field it goes, the longer it takes the photon to get out.

What is unexpected here is that the proper time goes to zero at the Schwarzschild radius. This means, effectively, that if a clock at that radius emits a photon, the photon will never get out!

How can this be? Recall that photons move in such a way that the spacetime-interval along their world lines is zero. Look at Equation 21.1 on page 285. If we are at $r = 2GM/c^2$, then the coefficient of $(\Delta t)^2$ vanishes. So a world line which remains at one spatial location,[†] i.e. with $\Delta r = \Delta\theta = \Delta\phi = 0$, has zero spacetime-interval, regardless of Δt . Therefore a photon can just sit on the surface $r = 2R_g$ forever, never getting out.

It follows that any photon emitted from *inside* $2R_g$ is trapped: not only will it not cross $2R_g$, it will in fact be pulled inwards, to smaller and smaller radii. Because light cannot come to us from inside $2R_g$ physicists call this surface the **horizon** of the spacetime.

[†]You might worry that at $r = R_g$ the coefficient of $(\Delta r)^2$ goes to infinity. To get around that, do this for r slightly larger than R_g and let r get smaller, always with Δr strictly equal to zero.

▷Remember how coordinate time was defined, illustrated in Figure 18.4 on page 230.

▷Notice that the Schwarzschild radius is exactly the same radius that Michell and Laplace had identified as the place from which light could not escape, as in Equation 4.12 on page 36. The main difference between the Schwarzschild black hole and the older Michell–Laplace black hole is that in relativity, photons never cross $2R_g$ from the inside. For Michell and Laplace, photons were like balls that were shot outwards at a fixed speed. If a ball started from inside $r = 2R_g$, it would simply move outwards and then fall back.

The gravitational redshift

In this section: we calculate the gravitational redshift in the full Schwarzschild geometry.



Figure 21.2. John Wheeler, one of the twentieth century's most imaginative and influential theoretical physicists. His many PhD students include Feynman (Chapter 19) and Thorne (Chapter 22). Wheeler was one of the physicists who helped clarify general relativity by focussing on what experimenters can measure. In this he followed the discipline of quantum physics, to which he had previously made key contributions. Wheeler coined the phrase "black hole". Photo by K Thorne, reprinted with permission.

In this section: the horizon separates the region that can communicate with far-distant astronomers from that which cannot. It is not a special place locally: an unwary astronaut could cross it and not know it until he found it impossible to get away.

In Investigation 2.2 on page 16 we found that a photon climbing a distance h in a gravitational field whose acceleration was g would have a lower frequency at the top than when it started, by the ratio

$$\frac{f_{\text{top}}}{f_{\text{bottom}}} = 1 - \frac{gh}{c^2}. \quad (21.3)$$

This is adequate for small distances h near the Earth, but in general relativity we need a more general form. We can deduce this for the Schwarzschild geometry from the clock equation, Equation 21.2 on page 286. Let us examine the relationship between the slowing of clocks and the gravitational redshift in the Schwarzschild geometry.

Suppose the clock near the place the photon comes from, at rest in the coordinates at radius r_0 , ticks once for each cycle of the photon, and the time between ticks on the clock is τ . Then the photon's frequency near the clock will be $f_0 = 1/\Delta\tau$. When the photon arrives at an experimenter far from the hole, where spacetime is effectively flat and the coordinate time t is proper time on clocks, then the distant observer can use the cycles of the photon to measure the ticking rate of the clock: the clock ticks once for every cycle of the photon. So the frequency f_{far} of the photon far away is the frequency of the clock ticks as measured by a distant experimenter. This redshift is, for our case,

$$\frac{f_{\text{far}}}{f_0} = \left(1 - \frac{2GM}{r_0c^2}\right)^{1/2}. \quad (21.4)$$

Notice that as the starting radius r_0 gets near $2R_g$, the frequency far away goes to zero. As we have seen earlier, these photons never leave the Schwarzschild radius at all.

Danger: horizon!

Now, are we in danger of a contradiction? Have we not learned in Chapter 15 that a photon cannot stand still? So how can a photon emitted at the horizon simply stay there? The answer is that the curvature of spacetime only makes the photon stand still with respect to an observer *far away*. Remember that general relativity incorporates special relativity only *locally*: the physics as observed by a freely-falling experimenter, looking only over a small region of spacetime, must be the same as in special relativity. If a freely-falling experimenter were to cross the horizon and observe the photon, it would not be standing still with respect to the experimenter: like any photon, it would be traveling with the speed of light relative to a local freely-falling observer. But time curves so much between the Schwarzschild radius and the distant observer that a photon traveling radially outwards at the speed of light only just manages to keep its distance from him.

Let us think about the freely-falling experimenter that we have just mentioned. Suppose she has taken the precaution of falling while strapped into a rocket ship with powerful motors. Her plan is to fall freely for a while, doing some important physics experiments, and then turn on the rocket and get away from the black hole. If she falls across the horizon before turning on the rockets, it will be too late: since nothing can move faster than light, not even a powerful rocket, and since inside the horizon *all* photons, no matter what their initial direction, fall inwards, the experimenter will also inexorably fall inwards. We shall consider what terrible things will happen to her in a later section.

The falling experimenter's risk is made worse by the fact that spacetime contains no signposts telling unwary travelers where the horizon is.

There is no sign saying, like the title of this section, “Danger: horizon!” Spacetime is smooth, empty, and locally flat there.

One can calculate the curvature of spacetime there, and it is unremarkable. There is curvature, of course, but it is not necessarily very large. In fact, the curvature of space and time at the horizon is proportional to $1/M^2$, so the larger the black hole, the smaller is the curvature at the horizon. The horizon only marks the boundary between where photons can get out and where they cannot, but that is a property of the large-scale structure of the spacetime, not something that can be sensed locally.

The nature of the horizon is illustrated by the light-cones drawn in the image under the text on page 285. They gradually tilt inwards as one goes towards the center, and the difference from one cone to its neighbor is small. Only by accumulating these small changes does the big difference between the cones far away and those at the horizon arise.

Getting away from it all

The slowing of time near a black hole has measurable effects besides the gravitational redshift. For example, it would in principle be possible to put something into a slow-time storage locker near a black hole. Suppose you don’t like your government, which is authoritarian and makes life in your country miserable, but you are not optimistic about outliving it. Just get into a rocket and spend a few days very close to a black hole. From the point of view of the government you left behind, your time near the black hole goes so slowly that your few days there take many years back home. When you come back the government may have changed and you will have only lost a few days. Of course, you may have lost all your family, friends and possessions, as well, but at least you are young enough to start over again!

There is nothing inconsistent in this. It certainly conflicts with our ordinary sense of how time behaves, but once we accept that gravity curves *time*, as we argued it must in Chapter 17, then such things become possible. This example may seem fanciful, but it is simply an exaggeration of what happens every day to the signals that go back and forth to the GPS satellites. And it is nearly duplicated by the gas orbiting black holes in galaxies: later in this chapter we will see the evidence for this.

Singularities, naked or otherwise

Let us return to consider what happens to the photon or observer that finds itself inside the horizon. Gravity forces it inwards, and here there is a real problem: at $r = 0$ there is what physicists call a **singularity** of spacetime: unlike at the horizon, the curvature of the spacetime does get infinitely large as one approaches $r = 0$, and so the tidal forces are arbitrarily large. Nothing could survive an encounter with such a singularity.

The existence of such a singularity is taken by many physicists to imply that general relativity cannot be a complete theory of gravity by itself: it does not predict what should happen to particles that encounter that singularity. There is good reason for not trusting general relativity near such singularities. They seem to imply the confinement of particles into very small regions, with very high energies. This may not be consistent with the Heisenberg uncertainty principle of quantum theory. It may be that when we have a consistent quantum theory of gravity, then we will find that the behavior of gravity near singularities is very different from what general relativity predicts.

Does this undermine our confidence in other aspects of black holes? Should we fear that quantum corrections will prevent black holes from forming in the centers of galaxies? The answer is no: the singularity is inside the horizon, and the horizon itself is not a place of strong curvature or any other effects that might make

▷Recall that the density of material forming a black hole also decreases with the mass of the hole. Large black holes exert locally rather weak gravitational forces.

In this section: how to use time dilation near a black hole.

▷Before signing up to a trip like this, you might think twice! Not only does everyone at home age much faster than you, but there are risks. You have to remain at rest very close to the horizon, which is not easy to see locally. If you make a tiny navigational error, you might wind up trying to park on the wrong side of the horizon, and your holiday would become your funeral!

In this section: the most disturbing aspect of black holes is that they contain singularities: places where general relativity breaks down. At least they are hidden inside the hole: if they were outside the hole (naked) then gravitation theory would be in trouble. The cosmic censorship hypothesis expresses physicists’ hope that this does not happen.

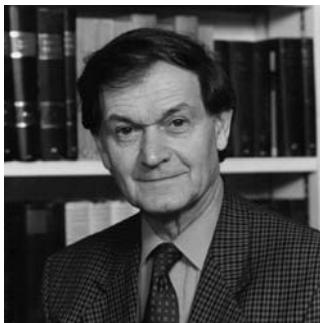


Figure 21.3. Roger Penrose has helped establish the modern mathematical framework for general relativity. Working with Hawking (see below) he showed that gravitational collapse generally produces singularities, which led him to the Cosmic Censorship Conjecture (see text). He has also stimulated a wide debate outside physics on the origins of consciousness. Photo courtesy Oxford University and Roger Penrose.

In this section: we modify the orbit program to compute orbits around black holes. There are new features, including orbits that are trapped by the hole. The motion of equal-mass objects under their mutual attraction can only be studied with supercomputers.

▷ Of course, just as in Newtonian gravity, once one goes inside the star the field will be different from that of a black hole, but outside they are identical.

quantum corrections important there. Most physicists believe that any changes to this picture that quantum gravity may bring will be confined to a region near $r = 0$, and will not even be visible from outside the hole. The horizon not only hides the singularity; it also hides our ignorance!

This point of view would be radically altered if it were possible to encounter singularities that are not hidden behind horizons. Such things are called **naked singularities**, and they would be a much more serious problem for general relativity than the singularities inside black holes. So far, no one has succeeded in establishing that very strong singularities can form naked. The contrary view is called the **cosmic censorship hypothesis**: framed by the British mathematical physicist Roger Penrose (b. 1931), it postulates that serious singularities will never be naked: Nature censors such sights from our eyes! The effort to prove (under suitably general conditions) or to find a counter-example to this hypothesis is one of the most interesting areas of mathematical research in relativity today.

The stakes are high. If naked singularities can form, then general relativity cannot be a complete theory: it cannot predict the future (even in its non-quantum form) even of systems that start out in a non-singular state. If this turns out to be true, then perhaps physicists will have to look toward a quantum theory of gravity (Chapter 27) to rescue the consistency of the entire theory, not just to make it compatible with the rest of physics!

What black holes can do ... to orbits

Let us turn once more to the effect of black holes on things near them, and specifically on the orbits of nearby particles. This is of more interest than just for black holes, since the gravitational field outside of a spherical star of mass M is identical to that of a black hole of the same mass. The study of orbits around black holes is then applicable to orbits around stars.

This is because, as we mentioned in Chapter 19, general relativity obeys a theorem similar to the one we proved for Newtonian gravity in Chapter 4: the gravitational field outside of a spherical body does not depend on the radius of the body, only on its mass, even if the radius is changing in time. Therefore, if a spherical star were to collapse in a perfectly spherical manner, in such a way that all its mass went to form a black hole, then the gravitational field outside the original radius of the star would not change at all. Planets outside would not even notice, provided the collapse was spherical or nearly so.

We saw this theorem in action earlier, where we observed that the gravitational field far outside a black hole is the same as that of a Newtonian star. Here we learn that this identity is true even close to the hole, where the field would be produced by a relativistic star of the kind we studied in Chapter 20. It turns out that, for neutron stars, the mass M of the star is the value of the quantity $m(r)$ at its surface. We already used this for the star's mass in plotting Figure 20.6 on page 281.

If orbits far from the hole are Newtonian, those nearby are certainly not. In Investigation 21.2 we develop the orbit equations for particles near black holes and relativistic stars. The modification from the Newtonian orbit equations is small, so it is an easy job to turn the program `Orbit` into `RelativisticOrbit`, which generates orbits around black holes. The details can be found on the website, and two example orbits are shown in Figure 21.4.

These orbits show important new features that are absent from Newtonian orbits. Both orbits begin close to the hole, at just five times the Schwarzschild radius, to insure that they show the effects without too much computer time. One of the orbits, which starts with a speed of $c/3$, spirals closer to the horizon and finally plunges across it. Such behavior is unknown in Newtonian gravity: no matter how near the

Investigation 21.2. Orbiting a black hole

In Investigation 4.1 on page 29 we developed the theory behind the program *Orbit*, which computed orbits in Newtonian gravity. Here we give the modifications that are necessary to do orbits around black holes and relativistic stars.

The acceleration of gravity produced by a Newtonian star of mass M has magnitude

$$a_{\text{Newtonian}} = \frac{GM}{r^2},$$

and it is directed towards the star. When we compute the acceleration components at an arbitrary location (x, y) (with the star at the origin), then we get Equations 4.5 and 4.6 on page 29:

$$a_x = -a_{\text{Newtonian}} \frac{x}{r} \quad \text{and} \quad a_y = -a_{\text{Newtonian}} \frac{y}{r}.$$

The only change that needs to be made to convert these equations into the ones that exactly describe motion around black holes is to change the magnitude of the acceleration $a_{\text{Newtonian}}$ to

$$a_{\text{relativistic}} = \frac{GM}{r^2} \left(1 + \frac{12K^2}{c^2 r^2} \right), \quad (21.5)$$

where K is the "Kepler constant" of the orbit, which is defined just as it was for Kepler's first law (see Chapter 4). That is, K is the area sweep rate of a Newtonian orbit with the same starting radius and velocity as the relativistic orbit has. As we remarked in Chapter 4, this rate is one-half of the angular momentum divided by the mass of the particle^a.

This extra term must be handled with care, because it involves the constant K that is not a property of the black hole but rather of the orbit. Once the starting position and velocity for the orbit are given,

^aExperts in relativity who have not seen this form of the orbit equations before may want to know more detail. The Cartesian coordinates used here are in the equatorial plane of the Schwarzschild solution, and are $x = r \cos \phi$, $y = r \sin \phi$, where r and ϕ are the usual Schwarzschild coordinates. All speeds and accelerations are changes in these coordinates with respect to the proper time of the orbiting particle. The angular momentum referred to is the usual conserved p_ϕ , so the Kepler constant is one-half of the specific angular momentum.

center an orbit gets, it will always come out again. But this particle is doomed in relativity by having too small an initial speed: it gets gobbled up by the hole.

The other orbit begins with 20% larger initial speed, and this is enough to keep it out of the hole. It follows a roughly elliptical orbit, but the orbit moves around. This is an extreme example of the precession we calculated in Chapter 17.

The plunge orbit illustrates the danger of the hole to particles near it. There are many orbits that start off innocuously, but which wind up being trapped. In fact, it can be shown that there are no stable circular orbits at all around a Schwarzschild black hole for particles nearer than three times the Schwarzschild radius! This radius is called the **innermost stable circular orbit**. You could try showing this with the orbit program. Start at, say, 2.5 times the Schwarzschild radius and vary the initial speed to see if you get any orbits that stay even roughly circular (such as a mildly elliptical precessing orbit). You will find none.

Although the acceleration of a single small particle near a black hole is a simple modification of the Newtonian acceleration, more complicated systems in relativity

one must calculate the area sweep rate and use it for K . The relativistic orbit equations given by this acceleration insure that this rate is constant, just as in Newtonian gravity.

Calculating the sweep rate for general starting values is difficult, so I have written the program *RelativisticOrbits* to start with values where it is simple. If the initial position is at a distance r along the x -axis and the initial velocity is v directed in the y -direction, then after a small time Δt the particle will have defined a right-angled triangle with one side of length r (the height of the triangle) and a perpendicular side of length $v \Delta t$ (the base). The area of the triangle is then $rv \Delta t/2$, and the area sweep rate is this divided by the time it took to generate the area, Δt . Therefore, the Kepler constant for this configuration is just $rv/2$.

Once calculated from the initial data, there is no need to recalculate the number: it remains the same. It just has to be used to compute the acceleration.

The last remark we need to make about our equations is the meaning of the time variable t when we use the given acceleration. It is the proper time of the orbiting particle, the time as kept on its own clock. So if you run the program and discover an orbital period, this is the time that elapses on the particle's own clocks during one orbit, not the time as seen by an observer watching the orbit from far away. To convert from one to another requires two steps: first use the special-relativistic time dilation to change from the particle's proper time to the time on clocks at rest along the particle's orbit, and then use the gravitational redshift given by Equation 21.4 on page 288 to convert to the time of the distant observer. This is not hard to do for an orbit that is perfectly circular, where the conversion factors are constant. It is rather more difficult for the orbits we have displayed.

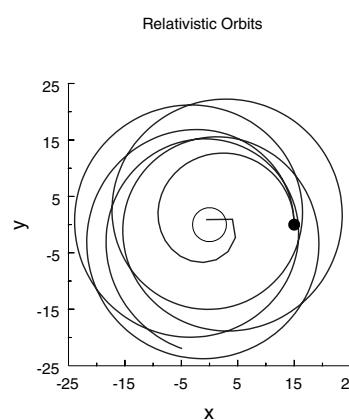


Figure 21.4. Two orbits near a black hole of one solar mass, as calculated by the program *RelativisticOrbits*. The horizon of the black hole is shown as the circle in the center. Both orbits begin at five times the horizon radius, shown by the heavy dot. One orbit does not have enough initial speed to stay away from the hole, and is captured. (The kink in the orbit at the end is due to the small number of points output by the computer program – the orbit is calculated accurately enough.) The other orbit stays near the hole but precesses by about half a circle on each orbit.

▷ In the terminology we used in Chapter 19, Einstein's equations are non-linear, so it is not possible to describe complex systems by adding together simpler ones.

are not generalizations of similar Newtonian systems. In relativity it is not possible to write down simple equations even for a binary system, as we did for Newtonian gravity in Investigation 13.1 on page 156. Even the geodesic solutions for orbits that we have just found are only approximations to what real particles will do near a black hole, because real particles will radiate gravitational waves (see the next chapter), lose energy, and gradually spiral towards the hole. The result is that realistic solutions for binaries or more complicated systems must be found on a computer.

Figure 21.5. This image illustrates the merger of two black holes of unequal mass, from an inspiralling orbit, as calculated in a supercomputer simulation. The outer surface is the horizon that forms around them both, indicating that by this stage of the simulation they have merged into a single black hole. The ghosts of the old individual horizons are still shown inside the new horizon, even though they do not have any significance at this point. Image adapted from one rendered by W Benger (ZIB) from a simulation performed by the AEI-WASHU-NCSA numerical relativity collaboration; used with permission.



ulation of the merging of two black holes.

Making a black hole: the bigger, the easier

In this section: the density of material just as it forms a black hole depends on the mass of the hole being formed. Holes smaller than a solar mass require densities higher than nuclear, which stops them being formed. Supermassive holes in galaxies can form from material less dense than water.

▷ The real volume of the gas cloud is not the volume of a sphere in Euclidean space, because space near the black hole is not flat. But we are not interested here in getting all the numbers exactly right, just in seeing approximately what numbers come out.

Several times in this book we have hinted that there are limits on what kinds of systems can remain as ordinary stars, and that if they cross those limits they will become black holes. General relativity alone does not tell us what those limits are, since the solution for the Schwarzschild black hole scales with the mass M , and this can have any arbitrary value. General relativity can contemplate black holes of 1 g or of one galaxy in mass, and is equally happy with both. The sizes of real black holes are determined by the astrophysical objects that make them.

One constraint on the physics of black hole formation is the density of material when it forms a black hole. A uniform cloud of gas of mass M just forming a black hole has a radius of $2R_g$. Taking its volume to be $4\pi(2R_g)^3/3$ then the density ρ_{bh} is roughly $M/(32\pi R_g^3/3)$. Putting in GM/c^2 for R_g gives, after some simplification,

$$\rho_{bh} \approx \frac{3c^6}{32\pi G^3 M^2} = 2 \times 10^{19} \left(\frac{M_\odot}{M} \right)^2 \text{ kg m}^{-3}. \quad (21.6)$$

The middle expression is not very informative, so I have evaluated it for the mass of the Sun, and shown how it scales from this value for any mass M of the black hole.

The density required to form a black hole scales inversely with the square of the mass of the black hole. Objects of small mass have to be compressed to high densities to form a black hole, while objects of very large mass can form black holes at low densities.

We can learn much from Equation 21.6. For example, if Nature were to try to form a black hole of the mass of the Sun, then the collapsing matter would have to

Such computer solutions are unfortunately not just simple generalizations of the computer work we do in this book. The solution of Einstein's equations must keep track of not just the positions of a few particles, but also the changing coefficients in the spacetime-interval everywhere. The whole of space and time is dynamical. Calculations like this require enormous computer memories and very fast processors. Supercomputers today (2002) are just becoming fast enough to do calculations like this, and scientists working in the field of numerical relativity hope soon to be able to provide physicists' first real insight into the behavior of strong gravitational fields in general relativity. Figure 21.5 shows an illustration of the output from a recent sim-

compress to a density of about $2 \times 10^{19} \text{ kg m}^{-3}$. This is a factor of 100 higher than the typical nuclear density of $2 \times 10^{17} \text{ kg m}^{-3}$, at which we saw in Chapter 20 the nuclear repulsion is strong enough to stop the collapse. We should not expect to see black holes of $1M_{\odot}$ in the Galaxy, and indeed none are known.

However, a black hole with a mass of $10M_{\odot}$ forms exactly at nuclear density. It is no surprise, then, that the black holes observed in X-ray binaries (as described later in this chapter) are typically this size.

The giant black holes in galaxies clearly form at low densities. Some are known with masses of $10^9 M_{\odot}$ or more. Such a black hole would form with a density of 20 kg m^{-3} , less than the density of water! We don't know if such objects formed in one grand collapse of a cloud of this density, or if they started out smaller and gobbled up smaller stars and black holes to reach their present size. But if they did form by a single collapse, then clearly nothing would have stopped the collapse: the density was not even high enough to have triggered nuclear reactions.

At the other extreme, black holes of small mass would be very dense when they formed. A black hole of 10^{12} kg , which we will see later is an interesting mass, would require a density of almost $10^{56} \text{ kg m}^{-3}$. A density this large was only ever seen once in the history of the Universe: shortly after the Big Bang. If such black holes exist, they could only have formed in the Big Bang. They are not forming now.

Inside the black hole

What about after the collapse? What happens to the material inside the black hole? Does it continue to collapse or does it stop when it finally reaches nuclear densities? The answer is that nothing can stop the collapse inside the hole, not even the hardcore nuclear repulsion.

The reason is that light cannot get out or even stand still in this region. Light is forced to fall inwards, even if we start it out in an outward direction. Therefore all freely-falling experimenters must also fall inwards.

If a collapsing star managed to stop collapsing and keep a fixed size inside the black hole, then it would be traveling outwards relative to all freely-falling experimenters at a speed faster than light. This is not possible. Once inside, the star must continue moving to smaller values of the radius r , reaching the singularity in a finite proper time.

So once matter has fallen in, it cannot stop collapsing. We will return later in this chapter to the natural next question, of what happens when it reaches the center.

Disturbed black holes

We have seen in Chapter 8 that a star has a natural vibration frequency that depends just on its density. Having calculated the density of a black hole, we can therefore compute its expected natural frequency from Equation 8.23 on page 100:

$$f_{bh} \approx \left(\frac{3}{32\pi} \right)^{1/2} \frac{c^3}{GM} = 35 \left(\frac{M}{1M_{\odot}} \right)^{-1} \text{ kHz.} \quad (21.7)$$

This is fairly close to the result of a full calculation of the ringing frequency of a black hole.

But how can a black hole oscillate? We have emphasized that the horizon is not a material surface, not a place which one can find by local experiments. Nevertheless, it does represent the boundary between what gets out and what is trapped, and the location of this boundary can change with time. If a small star falls into a black hole, it will disturb the location of this boundary, and the boundary will oscillate for a short time before settling down. Typically, it will execute only two or three

►We should be cautious because our estimate of the volume was rough and because gravity is very strong even before the collapsing cloud reaches the Schwarzschild radius. We saw in the last chapter that detailed calculations show that in fact gravity overwhelms nuclear physics when the star has a mass of no more than about $3M_{\odot}$, and perhaps even as low as $2M_{\odot}$. Our simple calculation here is consistent with this.

In this section: whatever falls inside the hole must continue moving to smaller radii. Nothing can hold itself up against such strong gravity. Everything reaches the singularity in a finite proper time.

In this section: black holes have natural frequencies of vibration, just as other objects do. These will give a characteristic imprint on gravitational radiation from events where holes are formed or collide.

oscillations in this way. The same happens when a black hole is formed by collapse, if the collapse is not perfectly spherical. The hole is formed in a non-spherical shape; the non-sphericity oscillates and dies away.

The oscillations produce a time-dependent gravitational field outside the hole, and as we will see in the next chapter this must lead to the emission of gravitational waves. So black holes can be dynamical and can emit gravitational radiation. Scientists are building detectors with the hope of finding this radiation, and numerical physicists are doing simulations to try to determine precisely the characteristics of this radiation. We will describe this more fully in the next chapter.

Limits on the possible

In this section: we look at some more simple numbers, the Einstein luminosity and the Planck scales, which lie at the boundaries of what we know about the physical world.

Our style of discussion here is called dimensional analysis. We met it first in Investigation 1.3 on page 5.

We have drawn some important conclusions in the last sections from rather simple calculations. Here we do some equally elementary calculations that take us to the limits of physics. Black holes, by their nature, fix boundaries between the possible and the impossible, and these boundaries depend on combinations of the most fundamental constants of Nature, c , G , and h .

We start with a simple observation, that it is possible to form a number with the dimensions of *luminosity* by using only c and G :

$$L_E = \frac{c^5}{G} = 3.63 \times 10^{52} \text{ W}. \quad (21.8)$$

This is called the **Einstein luminosity**. It is huge compared to the luminosity of most objects in the Universe. The luminosity of the Sun, for example, is only 3.8×10^{26} W, that of the Galaxy about 10^{40} W, and that of gamma-ray bursts around 10^{43} W for a few seconds. What significance can the very much larger Einstein luminosity have?

We will show here that L_E is an upper bound on all luminosities: anything that tried to have a larger luminosity would collapse to a black hole, pulled in by the self-gravity of the very energy it was trying to radiate away. To see this, consider the following extreme example. Suppose we have a system of mass M and radius R , which suddenly turns itself entirely into light, which then radiates away. We won't ask how this could happen; we will see whether a more realistic scenario teaches us anything after we follow this one through. So we have an energy of Mc^2 , and it leaves the region R at the speed of light, so it takes a time R/c to do it. That means the object briefly has the luminosity

$$L = Mc^2/(R/c) = Mc^3/R.$$

Now, the region's size can't be smaller than its Schwarzschild radius, or it would be a black hole and the radiation would not get out. So we can take $R \geq 2GM/c^2$. In turn this implies

$$L \leq c^5/2G = 0.5L_E.$$

More realistically, if the radiation escaping was only a small fraction f of the mass of the object, then the luminosity would be $0.5fL_E$.

The Einstein luminosity is the effective upper bound on the luminosity of any process.

Another universal number built out of fundamental constants is the Planck mass, which we met in Chapter 12. We shall see that we can also define the **Planck length** and **Planck time**. We reproduce Equation 12.20 on page 146 here:

$$m_{\text{Pl}} = (hc/G)^{1/2} = 5.5 \times 10^{-8} \text{ kg}. \quad (21.9)$$

Associated with this mass is its gravitational radius Gm_{Pl}/c^2 , called the Planck length:

$$r_{\text{Pl}} = (hG/c^3)^{1/2} = 4.0 \times 10^{-35} \text{ m.} \quad (21.10)$$

And then there is the time it takes light to travel the Planck distance, r_{Pl}/c , called the Planck time:

$$t_{\text{Pl}} = (hG/c^5)^{1/2} = 1.4 \times 10^{-43} \text{ s.} \quad (21.11)$$

These are all made only of fundamental constants. Since they include Planck's constant and Newton's constant, many physicists believe they must have something to do with quantum gravity. Perhaps r_{Pl} , for example, is the smallest length scale on which we could use general relativity to describe gravity.

To test this idea, let us introduce quantum ideas by using the Heisenberg uncertainty principle. Recall that a precise measurement of the size of something is accompanied by an uncertainty in its momentum, $\Delta x \Delta p \geq h$. Consider a black hole of mass M . For general relativity to be a valid description of Nature, it should be possible to localize the mass of the hole to within its gravitational radius, so that $\Delta x \leq GM/c^2$. It follows that the momentum of the hole cannot be defined to more precision than $\Delta p \geq hc^2/GM$. We can express this as a velocity uncertainty by dividing by the mass M to get $\Delta v \geq hc^2/GM^2$.

Now, for ordinary black holes this is really small. For a $10M_{\odot}$ black hole, the velocity uncertainty is $2 \times 10^{-67} \text{ m s}^{-1}$. We won't notice this! But if the velocity uncertainty is of the same order as the speed of light, then it will not be possible to talk about a black hole without using quantum theory, since it will be an object with no well-defined velocity or position. So if we put $\Delta v = c$ into the above expression and solve for M we get

$$M \geq m_{\text{Pl}}.$$

Physicists use the word **classical** to refer to non-quantum theories of physics, including general relativity. A classical black hole must have at least the Planck mass.

What might quantum black holes smaller than the Planck mass be like? Some scientists suggest that quantum fluctuations (see Chapter 7) might produce, temporarily, black holes of the Planck mass or smaller, which live for a time allowed by the uncertainty principle (h divided by their energy, $m_{\text{Pl}}c^2$, which gives the Planck time) and then disappear. This picture is called **spacetime foam**. The tiny black holes distort spacetime on the length-scale of the Planck length, so that it is not the smooth empty spacetime we think it is when we probe only on larger distance-scales. These are interesting ideas, but we will probably only know for sure what the Planck scales mean when we have a good theory of quantum gravity.

Other combinations of constants are possible. For example, the Planck density is m_{Pl} divided by r_{Pl}^3 ,

$$\rho_{\text{Pl}} = c^5/hG^2 = 8 \times 10^{95} \text{ kg m}^{-3}. \quad (21.12)$$

This is not a density we can think of achieving in laboratory experiments! But one would not trust a model of the Big Bang, for example, at times so early that the density of matter was higher than this: one would want a quantum theory of the Big Bang. We will take these issues up again in Chapter 27.

The uniqueness of the black hole

So far, we have treated only spherical black holes, as described by the Schwarzschild geometry. Since Schwarzschild was the first person to discover any exact solution of Einstein's equations, one might expect that further research would have revealed

In this section: black holes have mass, spin, and charge: that's all!

lots of other geometries describing black holes. Remarkably, this is not at all the case.

In fact, general relativity allows only one family of time-independent black hole solutions, of which the Schwarzschild geometry is one member (the one with zero spin and zero charge). In this family, the black hole is completely determined by giving only three numbers: its mass, its spin, and its electric charge.

When black holes are formed or disturbed, they can take on a variety of shapes temporarily. But after a few oscillations they settle down into a member of this single family by radiating away the disturbance in gravitational waves, as we saw earlier.

▷Imagine trying to describe everything about a *person* by giving just three measurements, for example the height, weight, and girth. It would be ridiculous to expect that any two people with the same height, weight, and girth would be identical: same hair color, same blood pressure, same sex, same dreams, same taste in food, same performance on a physics exam! It might seem even more ridiculous to expect this of black holes, which are formed by the collapse of enough matter to make at least 10^{25} human beings! Yet it is true.

▷Many physicists believe, however, that baryon and lepton number may not always be conserved. We will look at this in Chapter 25.

In this section: why spin shrinks the hole and makes a region outside it where nothing can stand still.

The idea that just three numbers fully describe a black hole is astounding. No other macroscopic object is as simple. So where has all the variety gone – the variety of all the stars with their sunspots and eruptions, the planets with their mountains and red spots, the countless individual grains of interstellar dust, whatever fell into the hole – that squashed together to form the black hole? It has, of course, disappeared into the hole. Once the material is inside the horizon, it can send no information back out, so this variety leaves *no trace* in the gravitational field outside the hole. The three numbers that remain are very special sums over what went in: the mass, spin, and charge of the hole as measured by an observer at rest with respect to the hole and far from it.

Why these three? Why not the total number of baryons, or some extra numbers that might describe the detailed shape of the horizon, a few extra ripples or corners?

Mass, spin, and charge characterize the black hole because they are the only three properties of a body that are conserved *and* that can be measured from far outside the body.

We learned in Chapter 19 that Einstein's equations demand the conservation of energy and momentum. For a body at rest, the linear momentum is zero but the spin (angular momentum) does not have to be. Both the total energy (mass) and angular momentum are conserved. Moreover, electric charge is also a conserved quantity. No chemical or nuclear reaction has ever been observed that changed the total electric charge.

Physics has other conservation laws. Nuclear reactions seem to preserve baryon and lepton number, for example. However, the nuclear forces are short range: we saw in Chapter 20 that they influence only other nearby baryons. Outside a nucleus one cannot measure directly how many baryons are in it, whereas by using the Lense–Thirring effect of gravity (Chapter 19) one could in principle measure the spin of the nucleus even from far away. Quantities conserved by short-range forces cannot be felt outside of a black hole: we will never know how many baryons fell into any given hole. Only three quantities are conserved and measurable from far away: mass, spin, and charge.

Forming a black hole results in a huge loss of information. In thermodynamics, this information is associated with a concept called *entropy*. We will see later in this chapter that the entropy of a black hole is immense.

Spinning black holes drag everything with them

The Schwarzschild black hole has no spin, but a realistic cloud of gas collapsing to a black hole should rotate, and should therefore form a spinning black hole. The solution of Einstein's equations that describes such a hole was not found until 1963,

by the New Zealand physicist Roy Kerr (b. 1934). The Kerr geometry is the unique family describing a time-independent black hole with spin and no electric charge.

The Kerr geometry brings gravitomagnetism to the black hole. Orbits that rotate in the same sense as the black hole experience the repulsion due to the gravitomagnetic effect that we calculated in Chapter 19. This allows them to approach closer to the hole, while remaining in circular orbits, than they could around a non-rotating hole. These closer orbits have shorter orbital periods. Astronomers believe they can now measure such orbits and verify that some known black holes are very rapidly rotating. (See Figure 21.7 on page 302.)

What is the horizon of a rotating black hole like? For the Schwarzschild black hole, we saw earlier that the horizon consists of light paths that stay at constant r , θ , and ϕ . For the Kerr horizon, the spin of the hole gives an extra repulsion to photons that orbit with the hole's rotation, so these are the ones best placed to resist the inward pull of gravity.

The horizon consists of photon world lines that rotate around the hole with the speed of the dragging of inertial frames on the horizon. The horizon is also smaller than the Schwarzschild horizon, since the repulsion allows some photons to come nearer the hole and still escape.

If the last photon to resist the inward pull of the black hole is rotating around it, then the light-cones near the hole are not only tilted inwards, as in the figure under the text on page 285, but also tilted in the direction of rotation. This must then be true for light-cones just outside the horizon, which means that photons and particles near the horizon must also rotate around the hole. Anything that stands still sufficiently close to the horizon must be following a world line that moves outside the light-cones, and is therefore going at faster than the speed of light relative to local freely-falling experimenters.

The horizon of a rotating black hole is surrounded by a region of finite size in which all particles and photons must move around the hole in the same direction as the hole rotates. It is impossible in this region to move backwards. The dragging of inertial frames near the horizon of a rotating black hole is irresistible.

The region in which dragging is so dominant has only a finite thickness. There is therefore another surface outside the horizon where it is possible for a photon to remain at rest with respect to a distant observer. This surface is called the **stationary limit**. Unlike the case for a non-rotating black hole, the stationary limit is not the horizon, because photons that rotate with the hole can escape from inside this surface.

The naked truth about fast black holes

Imagine letting particles fall into a black hole from orbits that have angular momentum in the same sense as the hole's. (This happens in black hole X-ray sources, as we will see below.) Then the hole's spin will increase, and the gravitomagnetic repulsion on co-rotating orbits will increase. By doing this it is possible to make the hole spin so fast that the gravitomagnetic repulsion is strong enough to allow photons to escape from any location: there is no longer a horizon!

The Kerr black hole that has just enough spin to annul the horizon is one whose angular momentum is related to its mass M by exactly $J = GM^2/c$. This is called the "extremal Kerr" black hole. However, the Kerr solution allows, at least mathematically, much larger values of J . In these, the horizon that normally conceals the inner

►It seems unlikely that astrophysical black holes will form with charge, so we will not describe the full family in this book.

►Remember we want always to describe geometry in terms of measurables. For the horizon, the measurable is its area. The radius of the horizon is just a coordinate, but the surface area is a geometrical quantity. The area of a spinning Kerr horizon is less than that of a Schwarzschild horizon of the same mass.

In this section: with enough spin, a Kerr hole can have a naked singularity. But most physicists believe that these holes cannot form.

singularity is gone, but the singularity remains: they are examples of *naked singularities*. Physicists have calculated that the repulsive effect of gravitomagnetism on material falling toward a Kerr black hole will prevent any real hole from gaining this much angular momentum, so that cosmic censorship will prevail. Observations of X-ray sources, as described below, can test this belief.

Mining the energy reservoir of a spinning black hole

In this section: the rotational energy of a spinning black hole can be extracted using negative-energy orbits outside the horizon. We show why these orbits exist and suggest how Nature is using them to power jets from quasars.

Like a massive spinning flywheel, the rotation of a spinning black hole represents a reservoir of usable energy. Physicists have learned in principle how to tap this reservoir, and it also appears that Nature has learned how to as well. It all has to do with the remarkable properties of particles and photons inside the stationary limit, where particles can have *negative* conserved energy.

We have met negative energy before, when we defined the gravitational potential energy of a particle in a Newtonian gravitational field in Equation 6.9 on page 54. It is part of the total energy of a particle on an orbit, the other part being its kinetic energy. It is negative because it is the energy given up to the kinetic energy as the particle falls. The sum of these two energies, the particle's total energy, is constant, and it equals the kinetic energy that the particle would have if it could reach a very distant experimenter.

The simplest particles to discuss in the Kerr geometry are photons. The conserved total energy of a photon is defined by analogy to the Newtonian case simply as the energy the photon has when it reaches a distant observer. In relativity the total energy includes any rest-mass the particle's may have.

For a photon, its *conserved* total energy is the gravitationally *redshifted* energy it has when it gets far away, regardless of what energy it began with.

Now, consider the photon that we mentioned earlier, which sits exactly on the stationary limit surface and is at rest with respect to a distant observer. A neighboring photon, just outside the stationary limit, would get out with a very large redshift, and therefore has very small positive total energy. The photon that is standing still does not get out at all, and therefore has *zero* total energy. It follows that a photon just inside the stationary limit could actually have *negative* total conserved energy. And if there are negative-energy photon world lines, then there must be particles with negative conserved energy inside the stationary limit, too, since particles can move as close as we like to the speed of light.

Of course, such negative-energy orbits are found inside the stationary limit in the Schwarzschild solution as well, but in the Schwarzschild case the stationary limit is the horizon, so these orbits are inside the black hole and of no relevance or use to us outside. However, some of these negative-energy orbits in the rotating Kerr geometry are outside the horizon, so they can participate in physical processes.

Particles in such negative-energy orbits can exchange energy with other particles. Penrose pointed out that this can be used to extract energy from the spinning black hole.

Imagine the following process. A distant astro-engineer drops two balls connected by a compressed spring toward the hole. Once inside the stationary limit, but still outside the horizon, the spring releases and the balls separate in such a way that one of them goes into an orbit that has negative total energy with respect to the distant engineer, although of course it has positive energy as measured by any local experimenter. Then, by conservation of energy (which still holds, even if some energies are negative), the other ball must be in an orbit whose total energy is more

- ▷ We will actually make use of the negative-energy photon orbits in the Schwarzschild geometry when we discuss the Hawking radiation below. The quantum uncertainty in their locations makes them accessible to some extent outside.
- ▷ Because of the negative-energy orbits, the stationary limit surface is sometimes called the **ergosphere**. This name uses the prefix *ergo-*, which indicates energy.

than the rest-mass energy of the original balls together. When this ball emerges from the stationary limit and reaches the distant engineer, it will have this large energy locally. The engineer has got back more energy than she put in: she is mining the hole!

The energy comes from the rotation of the hole, because the negative-energy orbits all have negative angular momentum, so that when the ball that is left behind inside the stationary limit falls across the horizon, the spin of the hole decreases. Reassuringly, there is no perpetual motion here: eventually the engineer will extract all the spin and the hole will become a useless Schwarzschild hole!

Notice that the engineer also extracts angular momentum from the hole. To mine the energy over a long period of time, she must also deal with the angular momentum.

Our astro-engineer is just a fantasy at present, but Nature may already be working the **Penrose process**. The giant black holes that power quasars and active galaxies are likely to be very rapidly rotating, because all the matter that they accrete comes in with high angular momentum. (We will look at the evidence for this in the next sections.) Moreover, as we saw in Chapter 14, these holes produce enormously powerful, highly collimated jets of gas, whose direction is sometimes stable over millions of years. More and more astronomers are beginning to believe that the source of the jets' energy is the rotation of the hole. It seems possible that magnetic fields generated in the accretion disk can penetrate the stationary limit and extract the rotational energy of the hole. Further research and many more observations will be required to discover the source of the energy in the jets.

Accretion onto black holes

Since most stars are members of binary systems, stars that form black holes usually start out in binaries. Unlike their cousins that form neutron stars, these events are much less likely to disrupt their binaries. Since much of the original star's mass stays in the black hole, the gravitational attraction between the binary stars does not weaken so much, and more of the systems will survive. So astronomers are not surprised to find that several percent of binary X-ray sources contain black holes.

As for the neutron star binaries we discussed in the previous chapter, it is often possible to estimate the mass of the accreting object from spectroscopic observations. Astronomers have studies systems that appear to contain black holes, which they call black hole candidates, to try to obtain at least a lower bound on the mass of the accreting object. When that lower bound exceeds, say, $5M_{\odot}$, then astronomers are confident that the object is not a neutron star: it must be a black hole.

The observations in fact usually indicate a mass around $10M_{\odot}$ for the compact object. This is why astronomers normally assume that black holes formed in stellar systems will typically have this mass. Table 21.1 gives a list of the best candidate black holes in stellar binary systems. Notice that they are all in our Galaxy or in the Magellanic Clouds. That is because X-ray telescopes are not yet sensitive enough to see many X-ray binaries in distant galaxies.

In Investigation 21.3 on the next page we look at the phenomenon of X-ray emission from accretion disks more closely, and learn why the compact objects at

Names	Estimated mass (M_{\odot})
GRO J0422+32 (XN PER 92)	≥ 9
A0620-00 (XN MON 75)	4.9–10
GRS 1124-683 (XN MUS 91)	5–7.5
4U1543-47	1.2–7.9
GRO J1655-40 (XN SCO 94)	7.02 ± 0.22
H1705-250 (XN OPH 77)	4.9 ± 1.3
GS 2000+250 (XN VUL 88)	8.5 ± 1.5
GS 2023+338 (V404 CYG)	12.3 ± 0.3
0538-641 (LMC X-3)	7–14
1956+350 (CYG X-1)	7–20

In this section: some X-ray sources contain accreting black holes. Although the main features of the X-rays are the same as for neutron stars, there are distinctive features that provide strong evidence that there is a hole in the center.

Table 21.1. Black hole candidates in stellar systems. The first column contains two names, the first being the modern style (catalog name, position on the sky) and the second (in round brackets) being the older name by which the object was known, often as a variable star or nova. The second column is the mass with uncertainties, or the mass range allowed by the observations. Data assembled by K Menou, E Quataert, and R Narayan (1998).

Investigation 21.3. X-rays from gas near black holes

When X-ray telescopes above the atmosphere began making observations at photon energies of 1–10 keV, they discovered a host of sources that had not been known at visible wavelengths. To emit substantially at these energies, the temperature must be of the order of this energy divided by Boltzmann's constant k , which gives a temperature of about 10^7 K. Compared to the surface temperature of the Sun (perhaps 5000 K) or of giant stars (cooler still), this is very hot.

When observations revealed that these sources had companion stars, then from optical observations of the stars a distance could be estimated. This in turn allowed estimates of the total luminosity

in X-rays. Typical values (for CYG X-1, for example) are 10^{30} W.

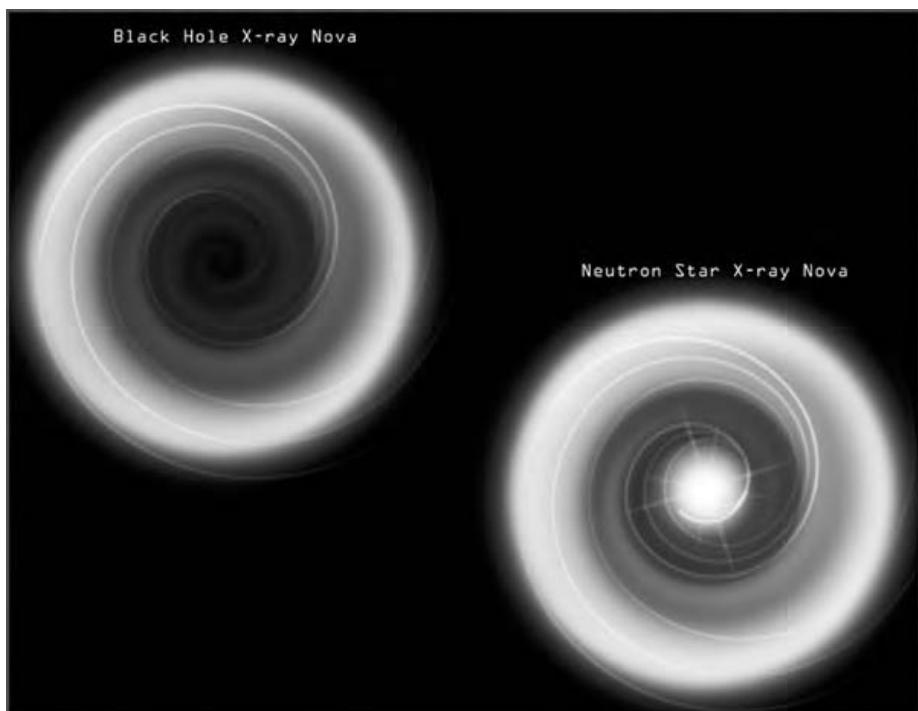
We studied accretion disks in Investigation 13.4 on page 161. Equation 13.9 on page 161 gives the luminosity, and by putting a temperature of 10^7 K into its right-hand side we can estimate that the area of the emitting region is that of a disk of radius only 24 km. This is an extraordinarily small region to be visible when it is so far away! Using Equation 13.8 on page 161 for mass falling onto a black hole of mass $15M_\odot$ (as in Table 21.1 on the previous page), we find an accretion rate of only 10^{-9} solar masses per year. Clearly this kind of source could last for a very long time.

Exercise 21.3.1: Accretion disks

- (a) If the spectrum of an X-ray source looks like a black-body spectrum that peaks around 1 keV, show that the associated temperature of the body should be near 10^7 K. (b) If the luminosity of the X-ray source is 10^{30} W, estimate the surface area and effective radius of the region emitting the X-rays. (c) Find the rate at which mass accretes onto the compact object, assuming that its mass is $15M_\odot$. Express the result in units of solar masses per year.

Figure 21.6. This drawing illustrates how astronomers can tell whether there is a black hole or a neutron star at the center of an accretion disk. If the spectrum of the radiation reveals a particularly hot component, it is likely to be coming from the surface of the neutron star, which is heated by the impact of the accreting matter. If that component is not there, then it is likely that there is no surface at all: the gas is falling across the horizon. The figure refers to an X-ray nova, which is like its white dwarf counterpart discussed in Chapter 13, except that the mass overflows onto a compact object.

These systems are ideal for the observations needed here, since it is possible to compare them during outbursts and quiet periods. Figure courtesy NASA/CXC/M. Weiss.



the centers of these disks must be either neutron stars or black holes.

The signature of the supermassive black hole in MCG-6-30-15

In this section: We look at the evidence that one particular galaxy has a massive black hole in the center. In this case, it is possible to use the spectrum of X-rays to show that there is gas within a radius that is only 1.3 times the gravitational radius. Not only does this establish that the object is a black hole, but it also shows that the hole must be very rapidly rotating.

Most black holes are known because they accrete. The argument that they are actually black holes is usually somewhat indirect. For massive black holes, the argument is often that the object is so compact that it can't be anything but a black hole. If they are of modest mass, they can be distinguished from neutron stars by looking for radiation from the surface of a neutron star, as in Figure 21.6. Increasingly, however, astronomical instruments are getting to be so good that better diagnostics are becoming available. It is even becoming possible to measure the spin of the black hole. As an example, we look here at how astronomers have detected the spin of the supermassive black hole in the galaxy called MCG-6-30-15.

This galaxy has an active nucleus, a kind of mini-quasar, and astronomers have long believed it to contain a massive black hole. One argument for a black hole is that the X-ray emission is variable on very short time-scales. Recently, the XMM-Newton satellite took an X-ray spectrum of this object. An X-ray spectrum is just like a spectrum in visible light: the X-rays are sorted by wavelength, and the spectrum records the intensity of X-radiation over a range of X-ray wavelengths. The spectrum of MCG-6-30-15 is shown in Figure 21.7 on the next page. It looks very different from the spectrum of visible light from a star, such as Figure 10.3 on page 114.

The spectrum contains a number of emission lines, which are wavelengths where the intensity is higher in a small region than elsewhere. This is expected in an accretion disk. These lines arise in this case from ions of heavy elements that have been stripped by collisions of all their electrons, which is normal at the high temperatures in an accretion disk. Occasionally an electron is captured by such an ion and drops into the lowest-energy state, emitting an X-ray photon that carries away the released energy. If the captured electron had no kinetic energy before it was captured, then the emitted photon's energy will be exactly the energy required to ionize the atom. In practice, the electron has some extra kinetic energy, so the emitted photons have a spread of energies above this value. In turn, the emitted wavelengths are spread over a range below a fixed value.

The wavelengths just described are measured in the rest frame of the ion. We would expect to see (at least) the three following different modifications to the emission line when the radiation comes from an accretion disk around a black hole.

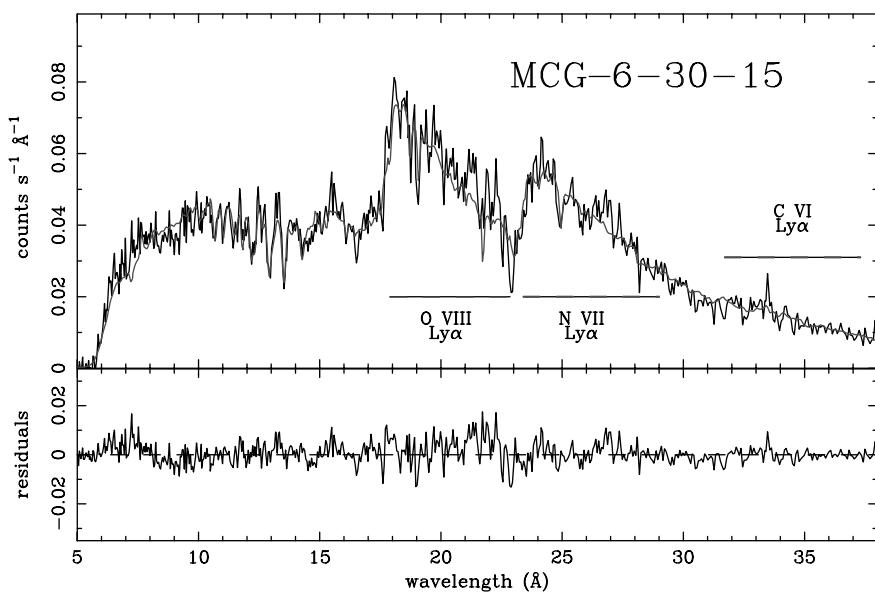
1. The lines should experience a gravitational redshift. In the spectrum of MCG-6-30-15, the lines are strongly redshifted. The redshift is the same for all the measured lines, so it is unlikely that the lines themselves have been misidentified.
2. The lines should be spread out more by the Doppler effect of thermal motion. The gas is hot, so the ions are moving a high random speeds. This is also seen in this spectrum, but not as strongly as the next effect.
3. The lines should be skewed because of the Doppler effect of the rotation of the accretion disk. If we are not looking straight down on the disk, then X-rays from the side of the disk in which material is moving away from us should be redshifted beyond the gravitational redshift, and those from the other side should be blueshifted relative to the gravitational redshift. Without relativity this would be a symmetrical effect, and would just broaden the line. But in relativity there is an effect called **beaming**. Radiation that is emitted isotropically in the rest frame of a particle will come out preferentially in the forward direction when it is moving. So the gas coming toward us in the accretion disk will emit more X-rays toward us than the gas moving away from us does. The result is that the line will be more intense at the shorter wavelengths.

In the case of MCG-6-30-15 the redshift is huge. For the line in Figure 21.7 on the next page labeled O VIII (an oxygen ion), the rest wavelength is about 14.2 Å, or 1.4×10^{-9} m. The center of the line is at about 2.1×10^{-9} m, and this should be roughly where the pure gravitational redshift can be estimated. That is a change by a factor of 1.5. If we use the relativistic redshift formula from the interval for the Schwarzschild black hole, Equation 21.4 on page 288, and remember that the ratio of wavelengths is the reciprocal of the ratio of the frequencies, we find that $(1 - 2GM/rc^2)^{-1/2} = 1.5$, which we can solve to find that $r = 3.6GM/c^2$. But we

>The interpretation of the X-ray spectrum of MCG-6-30-15 is an excellent example of modern astronomy at work. It shows how astronomers make use of both observations and theory (which they call modeling) to draw conclusions about such exotic objects. It also shows the key role played by technology: these observations were not possible before the satellite that made them was launched. Each new satellite and each advance in ground-based telescopes brings more data and leads to a more secure understanding of the Universe.

Figure 21.7. X-ray spectrum of radiation from the black hole in the center of the galaxy known as MCG-6-30-15, as measured by the satellite XMM-Newton. The dark jagged line represents the measured intensity of the X-rays (photons per second per unit wavelength interval) at the wavelength shown on the horizontal axis (in units of Å, or 10^{-10} m). The lighter line is a fit to a model in which the elements responsible for the broad features are fully ionized oxygen, nitrogen, and carbon. These data indicate that the inner edge of the accretion disk may be at $1.24GM/c^2$. As we discuss in the text, this black hole must therefore be spinning very rapidly. Figure from a paper by G Branduardi-Raymont and collaborators given at the Johns Hopkins University Workshop on X-ray Accretion onto Black Holes, proceedings at the website <http://www.pha.jhu.edu/groups/astro/workshop2001>.

Used with permission of the authors.



have already seen that accretion disks around a Schwarzschild black hole cannot extend within the last stable orbit at $6GM/c^2$. It can only go further in if helped by the repulsive effect of magnetogravity. So even from a crude inspection of the spectrum, we see that the black hole must be rapidly rotating. We are dealing with a Kerr black hole.

►The astronomers have made a model of the accretion disk that fully uses the Kerr geometry, accounting for all three effects listed above for X-rays emitted from everywhere in the disk. The model even makes a correction for the curved orbits of the photons that reach us after leaving the disk.

The authors of the paper from which the spectrum shown in Figure 21.7 was taken have done a careful calculation of what radiation to expect from a disk around a black hole with any spin. Their best fit to the data is the smooth line in Figure 21.7. Their estimate is that the inner edge of the accretion disk is at just $1.24GM/c^2$!

This is an astonishingly small radius. If the hole were Schwarzschild, it would be inside the horizon! This black hole must be rotating almost as fast as the extremal Kerr hole. And it must be a black hole: we are, after all, seeing radiation from just outside the horizon.

Wormholes: space and time tubes

In this section: wormholes are present in black hole solutions in general relativity, but they close off faster than anyone can get through them. Physicists speculate on how to keep them open with negative energy, and how to use them for time travel. The object of the speculation is to test the limits of general relativity.

One of the most intriguing aspects of black holes is that there are black hole solutions in general relativity that appear to involve “bridges” to other places. These bridges are really tubes that connect one part of space to another, and have acquired the name **wormholes**. The simplest Schwarzschild black hole involves such a bridge, connecting our space to a totally different three-dimensional space. The character of the Schwarzschild wormhole is illustrated in Figure 21.8.

Unfortunately, this wormhole does not provide a way of communicating with or traveling to the other region of space to which it is connected. Any particle or photon falling into the black hole will reach the singularity at $r = 0$, not the other end of the wormhole. What happens is that the wormhole is dynamical, and it

pinches off before anything can get through. The only way to pass through it is to go faster than light.

There is an even more important reason not to get too excited by the Schwarzschild wormhole: it is not present inside black holes that form by gravitational collapse. The reason is that the Schwarzschild solution only describes the *exterior* of the collapsing gas, and the wormhole is a feature of the interior. Inside the gas, there is no reason to expect connections to other parts of the universe to form spontaneously. They are only present in the mathematical solution that describes a black hole that is not formed by collapsing gas, but has existed forever.

Nevertheless, some scientists today are indeed excited by the prospects that wormholes may have some reality. The reason is quantum theory. It appears to be possible to keep a wormhole open long enough to allow a particle to pass through it if one can make *locally negative energy*. No ordinary physical systems have *locally* negative total energy – except in quantum theory.

Quantum theory allows uncertainties and fluctuations that are not allowed in non-quantum physics. Temporary fluctuations can produce photons of negative energy. In order to preserve the total energy, negative-energy photons form in pairs with positive-energy partners. These pairs almost immediately re-combine and disappear, since the quantum theory has to get rid of the negative-energy photon quickly in order to produce macroscopic physics of positive energy. But negative energy does exist for short times, in these quantum fluctuations. Like Planck-mass black holes, local wormholes may be an ingredient of spacetime foam.

More excitingly, physicists speculate that it might be possible to manipulate negative energy to build macroscopic wormholes. If they are able to open up and sustain a wormhole for a short time, then maybe an intrepid astronaut would be able to zoom through it into a different region of space!

Space travel is in fact only the second-most attractive possibility associated with wormholes: **time travel** is the first! Suppose the two ends of the wormhole emerge into the same space, next to each other. By accelerating one of them away from the other, it is possible to use the time dilation of special relativity (Chapter 15) to arrange that a particle that falls into the wormhole would emerge from the other end much earlier, in time to come back to the starting point just as it is about to fall into the wormhole the first time.

Ideas like these were until a few years ago just the province of science fiction. But even though they are now in the realm of serious scientific study, the expectations that scientists have about the importance of the ideas is very different from those of science fiction writers.

Some scientists feel strongly that time travel is a logical contradiction, and will somehow be ruled out by the laws of physics. The possibilities offered by wormholes are not, after all, solutions of the laws of physics, since we don't yet have the correct law of quantum gravity. Instead, they are speculations about the shape that such solutions can take. By studying the features of physical theories that may exclude

▷Actually, the Schwarzschild solution without gas starts at an infinite time in the past with a **white hole**, a time-reversed black hole. This is another reason that the pure-vacuum solution is unphysical; after all, our Universe appears not to have existed for all time in the past!

▷Gravitational potential energy can be negative, but that is not a locally defined energy: it requires a reference to a distant observer. Any local freely-falling experimenter measuring the energy of a particle will get just its rest-mass and kinetic energy, which are positive.

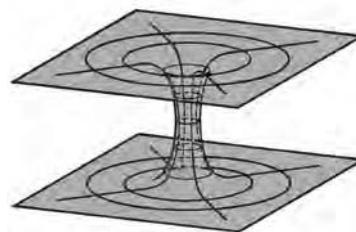


Figure 21.8. The wormhole of the Schwarzschild black hole. The upper sheet is a two-dimensional slice through our own three-dimensional space, and the lower sheet is another space, just as large, to which we are connected by the wormhole in the middle of the black hole. Unfortunately, this is only a picture of the situation at one moment. The wormhole actually gets smaller with time, and disappears completely before any particle or photon has time to get through. This wormhole is not a channel for communication.

In this section: remarkably, like other perfect absorbers, black holes actually radiate a black-body spectrum. This is a purely quantum effect. The radiation from astrophysical black holes is undetectably small, but in the early Universe small black holes might have formed that would explode today.



Figure 21.9. Stephen Hawking is one of the most well-known physicists of our time. He has been able to reach the general public with deep questions that are at the current limits of physical theory.

He has strongly influenced the development of theoretical physics, deepening the understanding of black holes and making a major step toward quantum gravity with his discovery black holes must emit thermal radiation. Photo courtesy S W Hawking.

► If you worry about our choice of wavelength here, consider that the uncertainty in the location of a photon is about one wavelength.

Very short wavelengths are localized outside the hole and fluctuations can't cross in time.

Very long wavelengths hardly notice the hole and have little chance of finding their way inside.

such phenomena, scientists hope to learn what some of the basic features of quantum gravity may be.

Hawking radiation: black holes are truly black bodies

Negative-energy fluctuations may be speculative when it comes to making wormholes, but they are well-established in another aspect of gravity: they turn black holes into black bodies.

Recall our discussion of black bodies in Chapter 10. There we saw that any body that absorbs all the light that falls on it is a black body, and when heated to a given temperature T it will give off a characteristic spectrum of radiation. Now, a black hole certainly absorbs all the light that falls on it, so it is a black body. But we have seen that nothing from inside can get out of a black hole, so it would appear that it cannot be a source of radiation. Black holes therefore don't seem to fit comfortably into thermal physics.

However, black-body radiation is a quantum phenomenon: Planck invented his constant in order to describe it. Fittingly, therefore, when the British physicist Stephen Hawking (b. 1942) studied the quantum theory of electromagnetism near black holes, he found that black holes actually emit radiation, that in fact has a black-body spectrum.

How can black holes emit radiation? It should be no surprise that the answer lies in quantum uncertainty. All over spacetime the quantum electromagnetic field is undergoing the little negative-energy quantum fluctuations that we considered above. Normally they are harmless and invisible, because the negative-energy photons disappear as quickly as they form. But near the horizon of a black hole, it is possible for such a photon to form outside the hole and cross into it.

Once inside, it is actually viable: as we remarked earlier, it is possible to find trajectories for photons inside the horizon that have negative total energy. So such a photon can just stay inside, and that leaves its positive-energy partner outside on its own. It has no choice but to continue moving outwards. It becomes one of the photons of the **Hawking radiation**.

In this picture, nothing actually crosses the horizon from inside to out. Instead, the negative-energy photon falls in, freeing the positive-energy photon. The net result of this is that the hole loses mass: the negative-energy photon makes a negative contribution to the mass of the hole when it goes in.

Once we accept that black holes can radiate, then it is not hard to estimate the wavelength of the radiation that they emit. The only length-scale in the problem is the size of the horizon. A photon with a wavelength λ equal to the radius of the black hole has (ignoring the curvature of spacetime in this simple argument) an energy equal to

$$E = h\nu = h\frac{c}{\lambda} = hc\frac{c^2}{2GM} = \frac{hc^3}{2GM}.$$

If black holes are indeed black bodies, absorbing everything that falls on them and emitting light, then their temperature T should be at least approximately related to this energy by setting $E = kT$, leading to the following estimate of the temperature of a black hole,

$$T = \frac{hc^3}{2kGM}.$$

Now, our argument cannot be expected to be exact, since we had no reason to take the wavelength equal to the radius of the hole rather than, say, its diameter or circumference, and since we must expect that the details of quantum theory and spacetime curvature will not be encapsulated in such a simple dimensional argument.

Investigation 21.4. The decay of a black hole

Here we study how long it takes a black hole to lose a significant amount of mass because of Hawking radiation. The temperature of a Schwarzschild black hole given in Equation 21.13 allows us to calculate the luminosity from the standard formula for a black body, Equation 10.3 on page 116,

$$L = \sigma AT^4,$$

where σ is the Stefan-Boltzmann constant, defined in Equation 10.4 on page 116, and A is the area of the surface that radiates. The surface in this case is the horizon, so the area is the area of a sphere with radius $2R_g$. (Recall that the radial coordinate that Schwarzschild used is the one that measures the circumference of the sphere, not the distance to its center. Therefore, we can be confident that the area of the sphere that is the horizon is given by the usual formula for spheres, even if we don't know what space is like inside the horizon.) This gives

$$A = 4\pi(2R_g)^2 = 4\pi(2GM/c^2)^2 = (16\pi G^2/c^4)M^2.$$

Combining this with all the other quantities gives the luminosity of the black hole, and grouping terms in a special way, gives

$$L_{bh} = \frac{1}{30720\pi^2} \frac{ch/G}{M^2} \frac{c^5}{G}.$$

It is instructive to take this expression apart. The first factor is, of course, just a pure number. The second contains, in its numerator,

the quantity ch/G . It is the square of the Planck mass m_{Pl} , defined in Equation 12.20 on page 146, which we have discussed elsewhere in this chapter. So the second factor is dimensionless, being the ratio of the squares of two masses. The third factor is the Einstein luminosity, also discussed in the body of this chapter.

The Einstein luminosity is large, but the black hole only approaches this luminosity when its mass is as small as the Planck mass. For an ordinary hole, the factor in $1/M^2$ reduces the luminosity drastically. For example, a $10M_\odot$ black hole radiates 10^{-30} W!

The lifetime of a black hole can be estimated to be Mc^2/L_{bh} ; this is an overestimate, since it assumes the luminosity will be constant in time, whereas it increases. But the increase is gradual, and so the estimate will be accurate to a factor of something like two. (A detailed calculation shows that the true lifetime is one-third of this estimate, which is not much error when we are dealing with such huge times.) For the $10M_\odot$ black hole, this estimate gives 2×10^{78} s, an unimaginably long time!

What is the mass of the hole that will just decay in the age of the Universe, about 10^{10} y, so that if these were formed in the early Universe, we would be seeing their explosions now? Just set the lifetime, Mc^2/L_{bh} , to this value and solve for M . The answer is that the hole should have a mass of about 10^{12} kg. This hole is too small to form today or at any time since galaxies formed, but perhaps in the very early universe conditions were different. There is no observational evidence for such holes, however.

Exercise 21.4.1: Hawking radiation

Perform the computations indicated in this investigation. Then find out how much time the hole has remaining when its temperature is high enough to produce electrons in its radiation (this will require kT to exceed $m_e c^2$).

Nevertheless, our answer is only a factor of $8\pi^2$ larger than the one that Hawking found, which is now called the **Hawking temperature** T_H :

$$T_H = \frac{hc^3}{16\pi^2 kGM} = 6 \times 10^{-8} \left(\frac{M}{M_\odot} \right)^{-1} \text{ K.} \quad (21.13)$$

This is so small for stellar-mass and supermassive black holes that it has little relevance to astrophysics. But Hawking's discovery is widely regarded as one of the first real steps toward a quantum theory of gravity. Although we have no such theory, many physicists expect that it must predict the Hawking radiation.

Through this radiation, black holes gradually lose mass. The smaller they get, the higher their temperature goes (by Equation 21.13), so the loss of mass accelerates. In Investigation 21.4 we use our knowledge about black-body radiation to calculate the lifetime of a black hole. For a one-solar-mass black hole, it is about 10^{67} y!

But smaller holes have shorter lifetimes. The mass of a hole that has a lifetime equal to the age of the Universe, about 10^{10} y, is 10^{12} kg. (See Investigation 21.4.) We have seen earlier that holes of this mass cannot form today, but it is conceivable that such **primordial black holes** did form by random fluctuations in the very early Universe.

These primordial black holes would be ending their lives today in an explosion. The amount of energy released in the last second of a black hole's life equals the energy equivalent of the mass of a black hole whose lifetime equals one second. This is a hole of mass about 10^6 kg, which converted into energy gives about 10^{23} J.

The Hawking radiation has linked black hole physics to two other, very different branches of physics: thermal physics and quantum gravity. When an unexpected result makes such links, they must be fundamental. In the next sections we will

▷ The release of this much energy in one second might be observable: it is only a fraction of a percent of the solar luminosity, but it would come out in gamma-rays; this does not explain the observed gamma-ray bursts. They have a luminosity that is up to 10^{22} times larger than this!

look at them further and learn why physicists find the Hawking radiation such a deeply satisfying result.

Black hole entropy: a link to nineteenth century physics

In this section: Hawking radiation allows physicists to define the entropy of a black hole. Entropy was introduced into thermal physics in the nineteenth century, and measures the disappearance of information. Since black holes swallow almost all the information that falls into them, they have extremely large entropy. Hawking radiation allows them to exchange entropy with other systems.

In this book we have discussed many aspects of gas dynamics in astronomy, but we have not yet studied the fundamental concept of entropy. Our study of black holes has led us to the point where it is now time to fill in this gap. The entropy of black holes is a remarkable illustration of the unity of the fundamental concepts of physics across different disciplines.

Entropy fundamentally has to do with measuring how much *information* a system contains. Information is related to order. An ordinary gas is highly disordered, its atoms moving in a random manner that is well described by only a few numbers, such as the density, composition, and temperature of the gas. A crystal lattice, on the other hand, has more structure, and correspondingly requires more information to describe it: the spatial arrangement of the atoms, their separations, the locations of any impurities, and so on. If a system is ordered, then it requires more information to describe it than if it is disordered. Entropy measures disorder. A highly ordered system has *low* entropy, and a messy system has *high* entropy.

Entropy was first introduced into gas dynamics by the German physicist Rudolf Clausius (1822–1888), but he did not associate it with disorder. This fundamental step was the greatest triumph of Boltzmann, whom we met in Chapter 7. He was able to show that his statistical mechanics, from which he could derive the pressure–density relation for gases, could also give a deeply satisfying definition of Clausius' entropy. Basically, Boltzmann showed that one could compute the entropy by counting the number of different ways that the molecules of a gas could be arranged to produce the same overall state of the gas: the same pressure, temperature, and density. This number is huge, of course, and the entropy is essentially the logarithm of it times the Boltzmann constant k .

Clausius had introduced entropy in order to describe heat flow. We have not needed to discuss it before because most of the fluid dynamics we have discussed in this book has been without heat conduction. In astronomy, heat flow is usually a secondary effect. But in engines and other technological systems, heat conduction is central to the function. Clausius originally defined the change in entropy of a system as the heat energy it absorbed divided by the temperature at which it absorbed the heat. When a system does things without losing heat, such as a gas expanding a piston in an idealized non-conducting environment, then the entropy of the gas did not change.

Since systems can gain or lose heat, their entropy can increase or decrease.

The remarkable discovery of Clausius was that – essentially because heat always moved from high-temperature regions to low-temperature ones – the total change in entropy, summed over all the parts of a system that were exchanging heat with one another, was always *positive*.

The entropy of the universe is always increasing.

This could be shown mathematically, but early physicists had no fundamental explanation for it.

Boltzmann showed that this was to do with disorder. Individual systems can get more ordered – I can clean up my desk once in a while, maybe – but the universe as a whole gets more disordered. When I clean my desk I expend so much energy that the entropy of the air in the room and of the chemicals in my body dramatically increase. (That's why I do it so rarely!)

It is one of the deep mysteries of the world that entropy increases, *disorder* increases, as time goes on. This so-called thermodynamic **arrow**

of time has intrigued physicists for a long time. The universe seems to be continually losing information.

As soon as physicists came to understand black holes, they realized that black holes have an interesting link to entropy. Black holes swallow up lots of information. They are universal wastebaskets. Since they refuse to tell us what has fallen in, they are systems which have the same external state for lots and lots of possible internal states. This suggests that they may have a large entropy. But how big is it?

The first step toward a measure of entropy was a theorem by Hawking that the area of a black hole must always increase, provided energy is always positive. For a Kerr black hole, which is not spherical, the area of the horizon depends on both the mass and the angular momentum. Hawking showed that, even for Penrose-type processes, which extract mass from the hole, they do it in such a way that the area still increases. The Israeli physicist Jacob Bekenstein (b. 1947), then working in the USA, recognized that the area was a kind of entropy function. But was the entropy a function of the area? A multiple of it? Could it be exchanged with other entropies? And what about information and disorder? The Hawking area theorem gave physicists a hint of entropy at the level of Clausius: something had to increase with time, but what did it mean?

The answer came with Hawking's later discovery of the thermal radiation from the hole. This gave physicists a chance to calculate the entropy, since they could then use the classical physics result that the decrease in the energy of a hole through its Hawking radiation, divided by the Hawking temperature, was the decrease in its entropy. The result gave the remarkably simple result that the entropy is proportional to the horizon area A :

$$S_{\text{bh}} = \frac{\pi k c^3}{2 G \hbar} A. \quad (21.14)$$

This is a huge entropy compared to that of ordinary objects. When a gas falls into a black hole, we really lose all information about it, and the entropy goes up enormously. When a black hole radiates some energy back into the outside world, the radiation carries its own entropy, so there can indeed be an exchange of entropy between black holes and other physical systems.

The study of the temperature and entropy of black holes is called black hole thermodynamics. It is remarkable that such exotic macroscopic objects as black holes can fit into the microscopic physics of Boltzmann in such a direct way.

Why should the value of the entropy depend on Planck's constant \hbar , i.e. involve quantum physics? Hawking has offered an explanation: that in a classical gravitational collapse, an outside observer never sees anything cross the horizon because time slows down near the horizon. Classically everything could always be observed, so no information would be lost. However, in a quantum world the hanging material could not be observed forever. Photons are quantized, so that eventually the material falling into the hole would send out its last photon, and then the outside observer would really have lost the information. So information is only lost because of quantum effects.

Black hole entropy: a link to twenty-first century physics

Despite their satisfaction with the unification of black holes with other thermal systems, physicists know that they are still working at the level of Clausius, able to define the entropy of a black hole, but not yet able to describe the link between black hole entropy and information or disorder. Most physicists believe that this link will help them towards another unification, that of gravity and quantum theory.

▷Here is the let-out for the Hawking radiation, which he discovered some years after his area theorem. We saw above that black holes lose energy and therefore shrink in area because of negative-energy photons falling into them.

In this section: Hawking entropy is seen by most physicists as a key beacon on the obscure road to a quantum theory of gravity. We give a plausible derivation of it.

Let us write Equation 21.14 on the previous page in another way that is suggestive of how quantum gravity might define the entropy. We replace Planck's constant by the appropriate function of Planck length given in Equation 21.10 on page 295. Then the entropy equation can be re-written in the simple form

$$S_{\text{bh}}/k = \frac{\pi}{2} \frac{A}{m_{\text{Pl}}^2}. \quad (21.15)$$

The entropy of a black hole is proportional to the number of "Planck areas" m_{Pl}^2 that would cover the area of its horizon. The proportionality is almost unity ($\pi/2$), apart from the requisite factor of k .

Now, we saw that in Boltzmann's statistical mechanics, the entropy of a system is k times the logarithm of the number of different microscopic ways that the given macroscopic system can be constructed. Taking this as a starting point, we might look for a way to count the number of ways a black hole can be made, as a way of calculating its entropy independently of the Hawking radiation. Maybe this number is the number of ways things can fall into the hole, or maybe it is the number of microscopic (quantum) states that would look like the same macroscopic black hole.

To see how this might work, consider the following rather simple approach, along lines originally suggested by Wheeler, whom we met earlier in this chapter. Imagine that in some quantum description of a black hole, the horizon is composed of an ensemble of "gravitons". (We shall discuss gravitons in Chapter 27.) These are presumably massless particles that travel at the speed of light and stay on the horizon. Suppose that the horizon is actually made up of such particles. Each might have an energy comparable to the Hawking energy, i.e. proportional to $1/M$. To make up the black hole mass M , the number N of such particles must be proportional to M^2 , or in other words proportional to the area of the horizon.

The entropy of the hole could be the logarithm of the number of ways the horizon could be constructed from such particles. In quantum theory, the particles are not distinguishable from one another unless they have different spin states. Now, it turns out that gravitons can have two different spin states, which correspond to the two independent polarizations of classical gravitational waves that we will learn about in Chapter 22. The number of different ways to build the horizon would be roughly 2^N , with some correction to make the total spin equal to zero. The logarithm of this is proportional to N and hence to the area of the hole.

Physicists are trying to make arguments like these more exact, and to ground them better in quantum gravity. For most such scientists, the calculation of the Hawking temperature and entropy is one of the acid tests of any proposed quantum theory of gravity.

►Physicists' attempts to find a fundamental derivation of black hole entropy have met with some success recently in **string theory**, which is one of the strongest candidates today for a method of unifying gravitation theory and quantum theory. Many take this as evidence in support of string theory. This illustrates the guiding role of Hawking radiation in physics today.

This discussion has taken us to one of the frontiers of theoretical research in gravitation theory. We will take up these questions again in Chapter 27, but in order to discuss them adequately we need first examine three further frontiers of research: gravitational waves, gravitational lensing, and cosmology. These are introduced in the next three chapters.

Gravitational waves: gravity speaks

One of the most radical changes in the behavior of gravity in going from Newton's theory to Einstein's is that Einstein's gravity has waves. When two stars orbit one another in a binary system, the gravitational field they create is constantly changing, responding to the changes in the positions of the stars. In any theory of gravity that respects special relativity, the information about these changes cannot reach distant experimenters faster than light. In general relativity, these changes in gravity ripple outwards at exactly the speed of light.

These gravitational waves offer a new way of observing astronomical systems whose gravity is changing. They are an attractive form of radiation to observe, because they are not scattered or absorbed by dust or plasma between the radiating system and the Earth: as we saw in Chapter 1, gravity always gets through. Unfortunately, the weakness of gravity, which we also noted in Chapter 1, poses a severe problem. Gravitational waves affect laboratory equipment so little that only recently has it become possible to build instruments sensitive enough to register them.

In this chapter we will learn what gravitational waves are, why scientists are confident that general relativity describes them correctly, how they are emitted by astronomical bodies, and what efforts are underway to detect them.

In our tour of the Universe so far, we have repeatedly seen that gravity is the engine at the heart of things, the force that dominates all others and controls stars, galaxies, and (as we will see in the final four chapters) the Universe itself. But gravity does not normally show itself to us directly. Scientists know what it does only because they can observe the photons (and sometimes other particles) emitted by the systems controlled by gravity. From these photons they infer, sometimes after long chains of deduction, what may really be going on inside the systems.

So far, gravity has been a silent engine. Scientists have never yet directly measured gravity from systems outside the Solar System. When gravitational wave detection becomes part of astronomy, astronomers will record in their laboratories the changing gravitational fields produced by some very distant bodies. Gravity will no longer be silent. It will tell us its story directly. Gravity will speak to us.

At this point scientists can only guess what it will say. This chapter looks at the best guesses they make today.

Gravitational waves are inevitable

From our explanation that gravitational waves simply arise from the restriction that no influence can travel faster than light, it is clear that gravitational waves will be a feature of any relativistic theory of gravity. Different theories may differ in the details of the waves, but all theories will have them. In this gravity is not unusual. All physical systems sustain waves: water waves, sound waves, pressure and buoyancy waves in stars, electromagnetic waves, and so on.

In this chapter: we meet the dynamical part of gravity. Gravitational waves are generated by mass-energy motions, carry energy, and act transversely as they pass through matter. Binary systems, involving compact stars or black holes, are the most important sources of detectable waves. The first detections are likely to be made by interferometers now under construction. The low-frequency observing window will be opened after 2010 by the planned international space-based LISA detector.

►The drawing on this page is of the LISA detector, which is described later in this chapter. LISA is being prepared by ESA and NASA for launch into an independent orbit around the Sun in 2011. It consists of three independent spacecraft using laser beams to track the changes in their separations. It will observe low-frequency gravitational waves from massive black holes in the centers of galaxies and from binary systems in our own Galaxy. From an image provided through the courtesy of the Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California.

In this section: special relativity forces any theory of gravity to have waves, but Laplace speculated about them two hundred years ago.

In fact, it is Newtonian gravity that is strange in this respect, because it does not have waves: when two stars in a binary system move around, their gravitational fields change instantaneously everywhere. So even if an experimenter is millions of light-years away, she could in principle feel the effect of the changing positions immediately, without any delay. This was called action at a distance, and some of Newton's contemporaries found this aspect of his theory disturbing.

Newton may have secretly shared this disquiet, but he was sensible enough not to let it deflect him from developing his theory. If he had tried to include wave effects, he could have hopelessly muddled the theory: experimental physics and astronomy in his day were simply not up to the job of measuring wave effects in gravity, and the whole theory might have been in trouble. Newton kept things as simple as he could.

The first physicist who seriously tried to work out the consequences of assuming that gravity might act with a delay and carry waves seems to have been Laplace. But his idea was that gravity was a kind of fluid, which emanated outwards from its source (such as the Sun) at a finite speed.

Laplace calculated that friction between the fluid and a planet would cause the planet's orbit to shrink. But he decided that, since no such shrinkage had been observed, the speed of gravity had to be large, in fact much larger than the speed of light. Laplace went no further with his speculations.

From the modern perspective, Laplace's speculations were impressive. He was on the right track: he wanted a finite speed and he looked for the right physical effect, orbital decay. What led him to the wrong conclusion is that he had the wrong model for gravity. The evidence, particularly electromagnetic theory and special relativity, that led Einstein to general relativity, was simply not available to Laplace. Given what he knew, he took a very modern point of view.

Knowing a little more than Laplace knew, we can already guess some of the properties of real gravitational waves. For example, in ordinary materials, the stiffer the material, the faster the wave speed. Since gravitational waves will travel with the fastest possible speed – the speed of light – it follows that space itself is effectively the stiffest possible “material”. In stiff materials, it takes a lot of force and energy to make a small disturbance, so we can expect that gravitational waves will have small amplitudes even when created by major events, like supernova explosions, and that they will carry large energies in their small amplitudes. We will see in this chapter how all of these guesses work out in Einstein's theory.

Transverse waves of tidal acceleration

In this section: gravitational waves act in the plane perpendicular to their direction of travel.

Just as the gravity of the Moon can be detected directly on the Earth through its time-dependent tidal forces (Chapter 5), so too are gravitational waves detectable through the time-dependent tidal accelerations they produce. The difference is that the force of the Moon comes from the curvature of time, whereas we will see that gravitational waves carry time-dependent spatial curvature.

The only part of the gravitational field of a wave that we can measure directly is the non-uniform part, which acts in such a way that one section of an apparatus is affected by gravity differently than another. We can therefore only register the *differences* in gravitational acceleration across the region occupied by our experiment.

In Chapter 5 we saw that the effect of the Moon on the Earth was to deform it from a sphere into an ellipse. Gravitational waves act in a similar way on objects they encounter, but relativity changes some of the details.

Gravitational waves produce tidal accelerations only in directions *perpendicular* to the direction they are traveling in. In general in physics, waves come with two types of action, producing motions either along or across the direction of motion.

Sound waves move air molecules back and forth in the same direction as the wave travels: this produces the compression and rarefaction of the air that constitutes sound. Physicists call sound waves **longitudinal**: they act *along* the wave direction. By contrast, waves on the surface of water are **transverse**: the water moves up and down as the wave moves across the surface. Electromagnetic waves are also transverse.

Gravitational waves similarly act transversely. We shall show that this is a consequence of the property of general relativity (inherited from Newtonian gravity) that spherical motions produce no changes in the gravitational field (Chapter 19). Suppose that a gravitational wave that encounters a slab of material acts on the material by producing alternating compression and rarefaction (by its tidal forces) along the direction of motion, just as does a sound wave. Then imagine taking a film of this and running it backwards in time. Any physical process run backwards also satisfies the basic equations of physics, so it is a possible (if unlikely) event. In this time-reversed film, the density of the material oscillates and *produces* gravitational waves.

This is a key concept: whatever action a gravitational wave has on matter is also the motion by which matter produces gravitational waves.

Now let us shape the material that produces waves into a perfect spherical shell. We arrange in some way that the shell oscillates in thickness, so that the density oscillates with time. Then this motion will produce gravitational waves, in our hypothesis. The waves must be spherical and will go outwards away from the shell, as well as inwards. But the mathematical theorem in general relativity that was quoted in Chapter 19 does not allow this: any spherical source produces a time-independent gravitational field outside it. Therefore, gravitational waves cannot act longitudinally. They must act transversely, like electromagnetic waves.

How gravitational waves act on matter

The analogy with electromagnetic waves breaks down when we consider the geometry of the way in which gravitational waves act on matter. Electromagnetic waves carry oscillating electric fields that make electrons move back and forth along a line, and the direction of the line is called the direction of the **polarization** of the wave. Gravitational waves are different. They produce deformations in the transverse plane that turn circles into ellipses, qualitatively similar to those we saw in Figure 5.2 on page 41. However, the deformation produced by the Moon is partly directed towards the Moon (the longitudinal direction), whereas gravitational waves are transverse.

Their action is illustrated in the top line of Figure 22.1 on the following page. The deformation ellipse that is produced by a wave has the same area as the original circle, so we say that gravitational waves in general relativity are *area-preserving*. Only two polarizations are illustrated, because only two are needed. The second is obtained by rotating the first by 45° . Any other action of a gravitational wave in the same plane can be described by combining these two.

Since the accelerations are *tidal*, the shape of the deformation is independent of the size of the original circle. If the circle were twice as large, the tidal accelerations across it would be twice as large, and the displacement these forces produce would be twice as large, leading to an ellipse with exactly the same shape, the same ratio of major to minor axes. Therefore, the measure of the strength of the gravitational

► You can prove that light is a transverse wave by using Polaroid, the semi-transparent material that is used in some sunglasses. If you take two pieces of Polaroid and place them over one another, then if they are oriented correctly they will pass about half the light through that falls on them. But if you rotate one piece by 90° , then the two pieces together will completely block all the light. This proves that light acts *across* its motion, because the rotation of one piece of Polaroid does not change anything in the longitudinal direction.

► The proof given here also works in electromagnetism to show that electromagnetic waves are transverse. This is because there is the same theorem for electric fields: a time-dependent spherical distribution of a given amount of electric charge has a static electric field outside it.

In this section: gravitational waves produce transverse tidal accelerations that deform circles into ellipses of the same area.

► To convince yourself that light acts along a line, look again at the experiment with Polaroid described in the margin earlier. When you rotate the second Polaroid sheet by 90° , then no light gets through. A further rotation by 90° restores transmission. The kind of geometrical object that is turned into itself by a 180° rotation is a line.

► In the case of light, there are also just two polarizations as well, but as the experiment shows they are obtained by rotating the line of action of light by 90° .

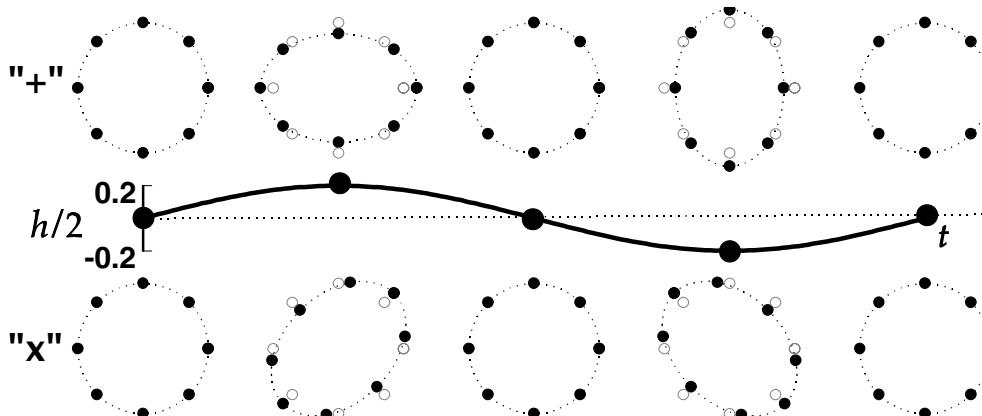


Figure 22.1. The two lines of circles show how gravitational waves act in general relativity by producing relative changes in proper distances between nearby free particles. They turn circles into ellipses of equal area, and there are two independent ways of doing this. These are the two different polarizations that a gravitational wave can have in general relativity, and they are labeled by their conventional names, "+" for the top line and "x" for the bottom; these represent the orientation of the axes of the ellipses. The middle line of the figure shows the deformation of the ellipse, which is half the gravitational wave amplitude h , so the axis is labeled $h/2$. It shows the wave as a function of time, and the ellipses top and bottom correspond to the times shown by the dots on the curve. The relative distortions shown are of the order of 20%. This is greatly exaggerated compared to what we expect from real waves, the strongest of which may produce relative deformations of order 10^{-21} . We will discuss the technology of achieving this below.

wave is this *relative* deformation. If the radius of the circle is called ℓ , and if the maximum displacement along an axis of the ellipse is called $\delta\ell$, then scientists define the *amplitude* h of the gravitational wave to be

$$h = 2 \frac{\delta\ell}{\ell}. \quad (22.1)$$

The factor of two is part of the definition, but we need not worry about why: this is just the definition that physicists have adopted. The relative deformation $\delta\ell/\ell$ itself is called the **strain** induced by the gravitational wave.

This figure shows the way the tidal accelerations produced by a gravitational wave deform a circle of particles if they are free to follow geodesics as the wave passes. (Remember of course that, because these are tidal accelerations, the whole assembly of particles may also have an overall free fall motion.) If the particles are part of a solid body, however, then the resulting deformation will be a result of all the forces, the tidal accelerations and the internal stresses of the material.

The fact that gravitational waves are transverse and do not act like the Moon does on the Earth implies that they are not part of the curvature of time, since that is where the Newtonian forces originate. They are purely a part of the curvature of space. When gravitational waves move through a region they do not induce differences between the rates of nearby clocks. Instead, they deform proper distances according to the pattern in Figure 22.1

Gravitational waves in other theories of gravity can act differently. Waves in theories called scalar theories of gravity are transverse but not area-preserving: the circle changes into a bigger or smaller circle. Some physicists expect that when general relativity is turned into a quantum theory of gravity (Chapter 27), it could become a scalar-tensor theory of gravity, in which case scalar gravitational waves might be present in real observations. So gravitational wave detectors will be trying to measure the pattern of action of any waves they detect.

Early confusion: are gravitational waves real?

The equivalence principle led to considerable misunderstanding and doubt about gravitational waves in the early development of general relativity. Physicists tended to think about waves acting like electromagnetic waves, which accelerate single particles relative to local inertial observers, and gravitational waves do not do that.

Many relativists, including at times Einstein himself, believed that gravitational waves were a mathematical illusion. The complexity of the mathematical formulation of general relativity prevented physicists from working out easily what the energy and momentum carried by the waves was. Some scientists believed that they could carry no energy, that gravitational waves were somehow not the same as waves in the rest of physics.

Fortunately, careful work by many physicists in the 1950s and 1960s clarified both the mathematics and the physics of gravitational waves. The picture I present in this chapter is the result of that work. Waves do carry energy, and we will see later in this chapter that astronomers observe the effects of the wave energy in certain astronomical systems. Just as with black holes, which were also not fully understood until the 1970s, astronomical observation has helped to clarify the physics of Einstein's equations.

I mention the early confusions here but will ignore them from now on. I have tried in the chapters on relativity and its consequences to distill the modern understanding of the theory down to the simplest principles and equations, to help you to see its logic and its physical content. But this understanding is the product of the work of dozens of the twentieth century's best physicists, who took Einstein's amazing baby and grew it to maturity. We are all today standing on the shoulders of those giants.

How gravitational waves are created

Imagine a gravitational wave emitted by a system somewhere, traveling through space, reaching the ring of particles drawn in Figure 22.1, and distorting them. Now imagine the whole process run backwards in time, as if you had taken a film of the wave and you now ran it backwards. As we have noted before, if something happens in physics, then its time-reverse is also possible, in the sense that it does not violate the laws of physics.

In the time-reversed film, the particles in Figure 22.1 move in and out in their elliptical pattern, the wave travels from the particles to what used to be its source, and the "source" moves in some way in response to the arrival of the wave. Now, what sort of motion is possible in the time-reversed "source"? Clearly, it can only move in the way shown in Figure 22.1, since it is responding to a gravitational wave in the time-reversed film. When we go back to the "real" situation, running forward in time, these motions are the ones that create the wave.

The kinds of motions that give rise to gravitational waves are similar to the motions in Figure 22.1. A source must deform in some kind of irregular way to emit radiation.

In particular, a spherical star that collapses but remains spherical only deforms circles into smaller circles, and this motion will emit *no* gravitational radiation in general relativity.

More particularly, here is how to judge whether and in what direction a source will radiate gravitational waves.

Look at the source from the desired direction. Since the waves act only in directions transverse to their motion, project the source's (perhaps

In this section: it took decades for physicists to cut through the mathematical complexity of general relativity and establish the physical reality of waves. They were helped by observations performed on the Hulse–Taylor binary pulsar PSR1913+16.

▷ This confusion is a good example of how physics develops. Although Einstein's theory emerged essentially complete in his 1915 papers, questions concerning the physical interpretation of the theory were not fully resolved until the 1970s. Physicists trying to discover the meaning of the theory were handicapped by its complex mathematics, and especially by the absence of any experiments or observations that could tell scientists how gravitational waves behaved. It was even difficult for physicists to understand just how confused they were: some physicists took passionate positions on the subject of gravitational waves, based on what we now know to have been flawed mathematical calculations.

In this section: we use a time-reversal argument to show that the motions in a detector mimic the motions of the source that creates the waves. Therefore the sources of waves are motions such as those in Figure 22.1.

▷ Not all physics has this time-reversibility property. It seems that certain reactions involving elementary particle do not conform: the time-reversal of some phenomena just do not occur. But in the macroscopic world of astronomy, all the theories of physics allow time-reversed behavior, even if it is wildly *improbable* (like the spontaneous heating of a cup of coffee, due to the cooling of its already cooler surroundings).

complicated) internal motions onto the "plane of the sky", which means onto a plane perpendicular to the line-of-sight to the source. Then only the motions in that plane that are some combination of the motions in Figure 22.1 on page 312 will generate gravitational waves. Moreover, the detector will respond with exactly the same combination of motions: detectors simply mimic the tidal distortions of the source.

We will use this rule when we come to the point in this chapter where we discuss radiation from various astrophysical sources.

The frequency of a gravitational wave is determined by the typical time-scale for things to happen in its source. If the masses radiating the waves move in and out, say, in 1 s, then the waves will have periods near one second and frequencies near 1 Hz. The upper bound on expected frequencies is about 10^4 Hz, because it is difficult to get large astronomical bodies, with masses comparable to the Sun or larger, to do anything on time-scales shorter than a tenth of millisecond or so. There is no lower bound, and in fact scientists are planning detectors that reach down below 10^{-3} Hz, also written as 1 mHz.

The set of all frequencies in a given gravitational wave is called its spectrum. Gravitational wave spectra are like sound spectra. There are some musical instruments that emit single notes, and others that emit thuds, bangs, or crashes. These instruments have counterparts in gravitational wave astronomy. Generally, a wave (either in sound or in gravity) will have a sharply defined frequency (its spectrum will contain a "line") if the motion of the source is regular and periodic or almost-periodic over a long time. Orbiting stars or stars vibrating in their normal modes (as in Chapter 8) emit narrow-line gravitational radiation. By contrast, a system that behaves in an irregular way, or whose radiation is so short-lived that there is time for perhaps one cycle of vibration or motion, emits a broad spectrum of waves, not concentrated sharply near any one frequency. The crashing gravitational collapse of the core of a giant star, which precedes a supernova explosion, could emit a broad spectrum of gravitational waves.

Strength of gravitational waves

In this section: we meet the quadrupole formula for the creation of gravitational waves.

For sources of gravitational waves that are not extremely relativistic, their Newtonian field effectively sets an upper limit on the amplitude of the emitted gravitational waves. If this were not the case, the tidal accelerations produced by the wave in its own source would exceed the source's self-gravitation, and tear the source apart; then we would have no source! For a source of mass M and size R , we should therefore expect that, as the wave leaves the source, $h \leq 2GM/Rc^2$, at least approximately.

Now, in all theories of gravity that have gravitational waves, the strength of the wave decreases as it moves away from the source, and this decrease is proportional to $1/r$, where r is the distance back to the source.

Thus, the wave amplitude remains a constant fraction of the overall distortion in the metric produced by the mass of the source.

Combining this with the limit above, we get

$$h \leq \frac{2GM}{rc^2} \quad \text{outside the source.} \quad (22.2)$$

This is already enough to tell us that realistic values of h should be small. For example, suppose we consider a gravitational wave coming from a neutron star in the Virgo Cluster of galaxies. Using a mass of $1.4M_\odot$ and a distance of 50 million

▷ This decrease in proportion to $1/r$ also happens to the amplitude of electromagnetic waves as they move away from a radiating antenna. We will see later that the energy flux carried by a wave is proportional to the *square* of its amplitude. The flux should therefore fall off as $1/r^2$, which is consistent with what we learned in Chapter 9 about light from a star.

Investigation 22.1. Energy flux of a gravitational wave

The flux of energy in a wave can be estimated from a physical argument. The only thing that a full calculation in general relativity is needed for is to get the coefficient in front right.

The energy flux must depend on the amplitude h of the wave, but it must not be simply proportional to h . Since h could be either positive or negative as part of the normal oscillation of the wave, anything that is proportional to h could also be positive to negative. The energy of the wave should not be affected by these regular oscillations. To make sure it carries positive energy, the flux should be proportional to h^2 . (It could be h^4 or any other even power of h , but we make the simplest guess.)

Next, the energy flux must depend on the frequency f of the wave. If it were independent of frequency, then a wave with zero frequency, which is just a value of h that is constant in time, would have the same energy as a wave of high frequency. But a zero-frequency "wave" should have zero energy, since nothing is changing, no energy is being transported. As with h , the flux must be proportional to an even power of f , since frequencies, like amplitudes, can be negative.^a Again for simplicity, we guess f^2 .

So far we have guessed $F = \alpha f^2 h^2$, where α is a proportionality constant that does not depend on the properties of the wave. What can it be? Surely it will contain some simple numbers, like 2 and π , but it can also contain some fundamental constants of physics. We can expect it to depend on c and G . But we should not expect other numbers, like Planck's constant or the mass of a proton to come into this, since such things are irrelevant to the energy carried by a pure gravitational wave in empty space.

Now we apply dimensional analysis. We know that the flux has units of energy per unit area per unit time, which is $\text{J m}^{-2} \text{s}^{-1}$. Since a joule is one $\text{kg m}^2 \text{s}^{-2}$, the dimensions of energy flux can be written as kg s^{-3} . Therefore αf^2 must have these units, since h is dimensionless. The dimension of the frequency f , since it is Hz, or oscillations per second, is s^{-1} . So we conclude that our unknown constant α must have dimensions kg s^{-1} . In particular it does not depend on meters.

These units for α must come from a combination of c and G . The dimensions of c are m s^{-1} , of $G \text{ m}^3 \text{kg}^{-1} \text{s}^2$. The only combination of

them that cancels the dimension of meters is c^3/G , or some power of it. It is easy to see that the dimensions of c^3/G are kg s^{-1} . This is exactly what we want for α ! So we have learned that α is a pure number (dimensionless) times c^3/G .

This is as far as our guessing method takes us. We have no way to guess the pure number. A full calculation in general relativity is required to get it right. The right value is $\pi/4$, which we include in Equation 22.4 on the following page. So even without a full calculation we came very close!

What have we learned from this analysis? We will answer that first by asking what could have gone wrong. Suppose we had not found a combination of c and G that gave the dimensions needed for α . If this had happened, then we would have had to go back to the beginning: we would have had to find a different dependence of F on f (possibly f^4 or f^{-2}), or we would have had to include some other physical property of the wave (but what is there besides its amplitude h and frequency f ?), or we would have had to include Planck's constant h (but that would have forced us to explain why defining energy in classical general relativity needs quantum theory).

Conversely, we might have erred by starting with a different guess for the dependence of F on f : we might have guessed $F = \alpha f^{-2} h^2$, for example. In this case, we would have stopped when we did the dimensional analysis: no combination of c and G would have given something with the units of energy flux. This would have been a clear signal to change the formula.

Altogether, we learn from this that, if gravitational waves carry energy, then the flux must be proportional to f^2 . Unfortunately, the dependence on h is not constrained by dimensional arguments. There we must rely on the reasonableness of our original assumption.

Ultimately, no guessing argument like this is fully satisfactory. Physicists have other, more deductive ways of defining energy, starting from the fundamental equations of general relativity. Fortunately for us, they give the same answer, and of course they pin down the dimensionless factor $\pi/4$. But for our purposes in this book, it is sufficient to be able to see that the expression for the flux that one gets from these more advanced methods is reasonable.

Exercise 22.1.1: Dimensional analysis

Fill in the missing steps above that show that the dimensions of energy flux are kg s^{-3} . Then show similarly that the dimensions of c^3/G times the square of the frequency are the same.

Exercise 22.1.2: Size of gravitational wave flux

We saw that a gravitational wave arriving at the Earth might have an amplitude h as large as 3×10^{-21} . If its frequency is 1000 Hz, then calculate the energy flux from such a wave. Compare this with the flux of energy in the light reaching us from a full Moon, $1.5 \times 10^{-3} \text{ W m}^{-2}$. Use Equation 9.2 on page 108 to compute the apparent magnitude of the source. Naturally, the source is not visible in light, so this magnitude does not mean a telescope could see it, but it gives an idea of how much energy is transported by the wave, compared to the energy we receive from other astronomical objects.

^aIf you are puzzled by the idea of a negative frequency, remember that frequency is the number of cycles of the wave per unit time. If we run time backwards, such as by making a film of the wave and running it backwards, then the number of cycles per unit time also goes backwards, and the wave has a negative frequency. But the backwards-running film shows a normal wave, one that you could have created in the forward direction of time with the right starting conditions, so it must also have a positive energy.

light-years, or $r = 4.6 \times 10^{23} \text{ m}$, we get $h \leq 6 \times 10^{-21}$. Our argument gives this as an *upper bound* on the strength of waves from such a source, and therefore on the distortions in shape that the wave produces in a detector.

How far below this upper bound do realistic wave amplitudes lie? Clearly this depends on the source. But when motions are not highly relativistic, it is possible in general relativity to make a simple approximation that works very well. The source must be the mass of the system radiating, since both the active gravitational mass and the active curvature mass are dominated by the ordinary mass-energy. But the overall mass of the system is constant and gives rise to the spherical Newtonian field, not to waves. We are looking for the part of the mass-energy that can follow

►This upper limit on realistic gravitational waves has set a target for detector developers since the 1960s.

the patterns in Figure 22.1 on page 312. It should not be surprising, therefore, that in general relativity:

gravitational radiation is produced only by the mass-equivalent of that part of the *kinetic energy* of the source that has the elliptical pattern of Figure 22.1 on page 312 as seen from the direction of the observer of the gravitational radiation.

Written as an equation, the prediction of general relativity is called the **quadrupole formula**. It is similar to the expression for the corrections to the coefficients of the interval that we computed in Chapter 18:

$$h = \frac{8G}{rc^2} \left(\frac{K}{c^2} \right)_{\text{projected elliptical part}} . \quad (22.3)$$

►The notation “projected elliptical part” here means that only the part of the kinetic energy that contributes to source motions similar to those of the test particles in Figure 22.1 on page 312 contributes to the radiation. Each polarization must be treated separately. The factor of eight is not something we can derive here; we must just accept that a full calculation in general relativity justifies it. It takes into account both the mass-energy and pressure parts of the source as well as any gravitational potential energy (the non-linearity of Einstein’s equations).

In this section: with the help of an analysis we calculate the energy carried by a gravitational wave. We see that even weak waves carry huge energies.

►We introduced the idea of energy flux in Chapter 9, where we discussed the apparent brightness of stars. The apparent magnitude of a star is a measure of the flux of light

energy we receive from it. By analogy, we have here the formula for the energy flux carried by a gravitational wave.

This gives a good approximation for radiation from systems where the velocities are small compared to c .

Einstein was the first to derive the quadrupole formula and yet, as I remarked earlier, he did not always have confidence in it. It took decades for physicists to be sure that it represented a good approximation, especially for realistic systems where gravitational potential energy was comparable with the kinetic energy. There were important contributions from Landau (whom we met in Chapter 20) and his Soviet colleague Yevgeny Lifshitz (1915–1985), and from Chandrasekhar (see Chapter 12), among many others. The subject is still an important area of research today, though not a controversial one. Physicists are developing better and better approximations to the radiation by refining Equation 22.3, in order to be able to recognize and interpret gravitational waves in the observations made by the detectors that we will describe below.

Gravitational waves carry energy, lots of energy

Gravitational waves clearly can transfer energy from one system to another. For example, if the particles in Figure 22.1 on page 312 are embedded in a viscous fluid, then their motion will transfer energy to the fluid, and long after the wave is gone the energy will remain. The energy transferred should be small, because we know the waves have great penetrating power.

To find out what energy is carried by the waves requires a small calculation, so it is reserved to Investigation 22.1 on the preceding page. The result, however, is important enough to write down here. Let us consider a **plane wave**. This is a wave from a source that is so far away that the wave passes us with a flat wave front, all parts of the wave traveling in the same direction with the same amplitude h . Suppose in addition that the gravitational wave is a simple sine-wave oscillation with a frequency f (measured in hertz). The appropriate measure of the energy carried by the wave is its *energy flux*, the energy carried by the wave through a unit area per unit time. The formula derived in Investigation 22.1 on the previous page is

$$F = \frac{\pi c^3}{4G} f^2 h^2 . \quad (22.4)$$

The key point about this formula is that the energy is proportional to the squares of the amplitude and of the frequency. Each of the two polarizations of the wave contributes its own energy, so this formula must be used separately for the “+” and “ \times ” amplitudes.

The constant c^3/G is a very large number, so that even when h is as small as we have found it to be, the flux can be large. In Exercise 22.1.2 on the previous page

we find that the flux of energy carried by a gravitational wave can easily be larger than the flux of light energy we receive from a full Moon. Considering that the source of the wave could be in the Virgo Cluster of galaxies, while the Moon is by comparison right next door, it is clear that the emission of gravitational radiation by an astronomical object can be a catastrophic event, carrying away huge amounts of energy.

Because the equivalence principle allows us to wipe out any gravitational field locally, even a gravitational wave, the energy of a wave is really only well-defined as an *average* over a region of space whose size is larger than the wavelength of the wave, and over a time longer than the period of the wave. Extended bodies can therefore only extract the energy if they interact with the wave over a long enough time or a large enough distance.

In the present case, the geometry of spacetime is constantly changing because of the gravitational wave, so energy conservation needs to be treated carefully. Indeed, if we consider just a matter system (such as a detector for gravitational waves), then the waves are an external time-dependent influence on it, and we do not expect its energy to be constant. That is good: one hopes a wave will disturb the detector enough to allow us to measure it! To arrive at a conserved energy that can be exchanged between the detector and the wave, we have to treat the wave and detector together. This is not so easy in general relativity, because it is not easy to define the wave separately from the rest of the geometry.

To see the reason for this, consider water waves. Drop a rubber duck into the still water in a bathtub. Waves ripple out from the place where it lands. We have no trouble distinguishing the waves from the rest of the water, and eventually the waves disappear and we return to the same still water surface as before. By contrast, look at a stormy ocean during, say, a hurricane. Near the beach, what are the waves? Sometimes there is water, sometimes beach. The whole ocean is moving. There is no way to define waves as a disturbance *on* the water.

Strong, time-dependent gravitational fields must be treated with more care in general relativity than we are able to do here. Recall that we learned in Chapter 6 that energy is only conserved in situations where external forces are independent of time. For weak waves, it is possible to define their energy with reference to the “background” or undisturbed geometry, which is there before the wave arrives and after it passes. But if the geometry is strongly distorted, the distinction between wave and background has little meaning. In such cases, physicists do not speak about waves. They only speak of the time-dependent geometry. But normally such regions are small, and outside of them the waves take shape as they move away.

The Binary Pulsar: a Nobel-Prize laboratory

In 1974, two astronomers made a discovery that was finally to give gravitational radiation theory an experimental foundation. The American radio astronomer Joseph H Taylor (b. 1941) had sent his graduate student, Russell Hulse (b. 1950), to observe pulsars with the largest radio telescope in the world, the Arecibo telescope in Puerto Rico. Hulse noticed a signal that appeared to be a pulsar, but strangely its pulse frequency kept changing. He told Taylor, who soon joined him at Arecibo, and together they determined that the pulsar was changing its frequency in a periodic way, coming back to its original frequency every eight hours or so. For a star like a neutron star to change its rotation speed that rapidly seemed impossible, like trying to slow a thundering train. Something else had to be making the frequency change. The conclusion was inescapable: the pulsar was in orbit around another star, with a period of eight hours, and the change in the frequency was simply the Doppler effect as the pulsar went away and came back again in its orbit.

>The weakness of the influence of the gravitational wave on the Earth shows that little of the energy carried by the wave is left in a detector. This is due to the weakness of gravity itself, not to any lack of energy in the waves.



Figure 22.2. Joseph Taylor.
(Photograph by Robert Matthew provided courtesy Princeton University.)



Figure 22.3. Russell Hulse.
(Photograph provided courtesy Princeton Plasma Physics Laboratory.)

In this section: the discovery of the first pulsar in a binary system provided the first experimental confirmation of the theory of gravitational radiation. It has become a test of extraordinary accuracy.

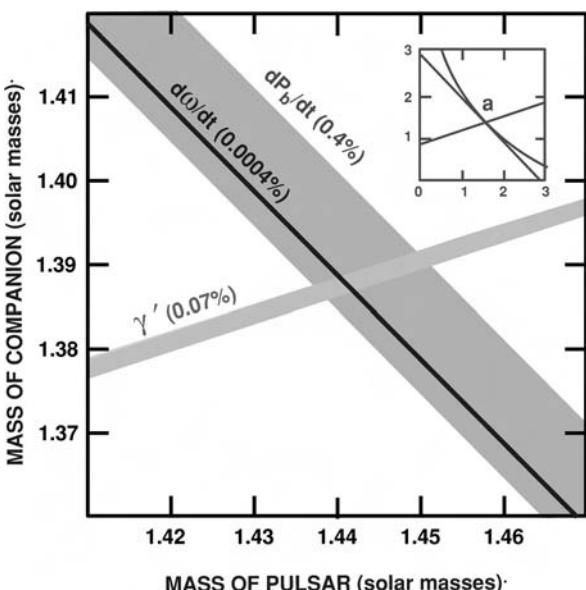
But an 8 h period is extraordinarily short. No binary had ever before been observed with such a short period. Mercury goes around the Sun in 88 days. A satellite of the Sun would have to be just skimming its surface in order to have an orbital period as small as 8 h. But the pulsar was not skimming a star: there was no evidence of friction making the orbit change quickly, and later optical observations of the pulsar's position did not reveal any star, not even a white dwarf. The pulsar, therefore, was orbiting another neutron star or a black hole. Whatever it is there radiates nothing we can see. Because this was the first pulsar discovered in a binary system, astronomers began to call it *the Binary Pulsar*. Radio astronomers have subsequently discovered many other pulsars in binaries, so the name is no longer a good one. We shall call it the Hulse–Taylor binary pulsar or simply PSR1913+16.

The orbit of PSR1913+16 is highly relativistic, its speed being about 0.1% of the speed of light. The orbit is, fortunately, a rather eccentric ellipse, so the precession of the perihelion (in this case, it is called the **periastron**) is easy to measure because it is 4° per year. (Compare this to Mercury, where one waits a century or so for the effect to build up enough to measure it accurately!) As we saw in Chapter 18, the precession depends on the mass of the companion, but when (as is the case here) the satellite's mass is not negligible compared to the companion, it is not possible to determine each mass individually from the precession alone.

But Taylor, by repeated careful observations spread over many months, was able to extract another relativistic effect. He could see the change in the pulsar's spin rate as it moved closer to and further from its companion. As we discussed in Chapter 20, this is caused by two effects that act together. The first is the changing gravitational redshift as the pulsar moves in and out in the companion's gravitational field; this redshift affects the spin rate in the same way it would any other clock. The second is the bending of the path of the radio waves as they pass near the companion, which introduces a changing time-delay that adds to the gravitational redshift. The combination of these two effects and the precession allowed Taylor to deduce the masses of both stars, as shown in Figure 22.4. Remarkably, they are both of mass about $1.4M_\odot$. Today the masses are known to an accuracy of better than 0.1%, the best mass determinations of any objects outside the Solar System.

Figure 22.4. This figure shows the way the masses of the stars in the Hulse–Taylor pulsar system are determined, and how the observed period decrease is consistent with them. The axes are the masses of the two stars, and the lines show how the observed properties of the system depend on the masses. The line labeled γ' is the combined redshift and time-delay term. Any combination of stellar masses on this line would give the observed delay. The width of the line indicates the spread of values allowed by the observations. The extremely narrow line labeled $d\omega/dt$ is the region allowed by measurements of the periastron shift of the elliptical orbit. The narrowness of this line shows how well this is determined. The broader area around this line, labeled dP_b/dt , is

the region allowed by the observed shortening of the orbital period. The fact that all three strips overlap in one region (at masses about 1.39 and 1.44 times the mass of the Sun) is a strong test of general relativity. In another theory of gravity, they need not coincide. The inset figure is the same figure drawn with a larger range of masses. This shows that the curve for the orbital period bends away from the periastron curve over a larger region; if general relativity were not correct then these two curves might not touch at all. Figure courtesy of C M Will.



Because the companion has a mass in the range of masses of neutron stars, it seems unlikely it could be a black hole: pressure would have halted its collapse. So it is assumed to be another neutron star, but there are no direct observations of it, no pulses of radiation or faint glow of X-rays that might confirm this.

The Hulse–Taylor pulsar is a laboratory for relativity. It confirms the perihelion precession calculated by Einstein to much higher accuracy than Mercury does. It demonstrates the gravitational redshift of a huge clock, showing that the equivalence principle works even for timekeeping by the spin of relativistic stars. All this information is enough to

Example system	Component mass M	Orbit radius R	Distance r	f_{gw} (Hz)	t_{gw}	h	L_{gw} (L_{\odot})
Hulse–Taylor	$1.4M_{\odot}$	1×10^6 km	8 kpc	6.9×10^{-5}	7.4×10^9 y	3.5×10^{-23}	1.5×10^{-3}
NS-NS	$1.4M_{\odot}$	50 km	200 Mpc	190	1.5 s	2.8×10^{-23}	4.7×10^{18}
MBH–MBH	$1.4 \times 10^6 M_{\odot}$	5×10^7 km	4 Gpc	1.9×10^{-4}	1.5×10^6 s	1.4×10^{-18}	4.7×10^{18}

Table 22.1. Three binary systems of the type that could be detected by ground-based or space-based gravitational wave detectors. For simplicity the systems are assumed to contain equal-mass components in a circular orbit around one another. For each example we specify the masses of the stars, the orbital radius, and the system's distance from us; then we calculate the frequency of the gravitational waves f_{gw} from Equation 22.6 on the following page, the chirp time t_{gw} (orbital shrinking time-scale due to gravitational waves) from Equation 22.12 on page 321, the maximum gravitational wave amplitude h at the Earth from Equation 22.7 on the following page, and the gravitational wave luminosity L_{gw} from Equation 22.10 on the next page. The latter is given in units of the solar luminosity L_{\odot} . For the system in the first line, which is a circular-orbit version of the Hulse–Taylor binary pulsar system, the calculated chirp time is longer than the observed one by a factor of about 12, because of the eccentricity of the real orbit. This brings the stars closer together for a fraction of their orbits, and so the average value of the luminosity is larger. The second and third systems are binaries that have the same compactness, as measured by GM/Rc^2 . Notice that they have the same luminosity, despite having very different masses. The more massive system (third line) has a longer lifetime, allowing it to radiate more energy in total. The third system also has the strongest amplitude despite being at a very great distance, where the cosmological expansion redshift is about one.

tell us everything we would want to know about the orbit.

And on top of all of this, the orbit shrinks. As gravitational waves carry energy away from the orbit, the stars get closer together, and the orbital period decreases. This is exactly the effect Laplace looked for in planetary orbits. General relativity of course provides a prediction for the rate of shrinking, and it has no adjustable numbers in it. Since physicists know the masses and separations of the stars from the other relativistic effects, they can use general relativity to predict exactly how rapidly the period should decrease. We make an estimate of the energy radiated by the system in Investigation 22.2 on the next page, and from it the expected rate of change of the period in Investigation 22.3 on page 321. The prediction is that the period should lose $(2.4427 \pm 0.00005) \times 10^{-12}$ seconds per second. The uncertainty of $\pm 0.00005 \times 10^{-12}$ seconds per second comes from the uncertainties in the deduced masses of the stars. The measurement is that the system is losing $(2.4349 \pm 0.010) \times 10^{-12}$ seconds per second. The uncertainty here is the observational accuracy. The two numbers agree within the uncertainties, as is shown in Figure 22.4.

This is a stringent test of general relativity and a striking confirmation of the predictions of the theory regarding gravitational radiation. For their discovery of this immensely important system, Hulse and Taylor received the Nobel Prize for Physics in 1993. Unlike the case of Jocelyn Bell, to which we referred in Chapter 20, in this case the Nobel committee included the graduate student who first recognized the phenomenon. Perhaps the controversy over Bell's omission was a lesson learned by that committee.

►The shrinking of the orbit happens because general relativity creates a small gravitational radiation reaction force, so named because it is the reaction of the orbit to the loss of energy to gravitational waves. We mentioned this in Chapter 2.

Gravitational waves from binary systems

Although the Hulse–Taylor binary system is radiating gravitational waves with a strength that physicists can compute exactly, there is little hope of directly detecting them in the near future: their frequency (given in Table 22.1) is too low for detectors now being planned, as we discuss later. Nevertheless, other binary systems are the most important gravitational wave sources that the detectors now planned or under construction will search for.

Astronomers now know that there are many other binaries with even shorter periods than the Hulse–Taylor system. A few systems that are known from optical or X-ray observations in our Galaxy have periods that will be detectable by the space-based detector LISA, which we will describe at the end of this chapter. Even

In this section: there is a wide variety of binary systems that could be radiating detectable gravitational waves. Coalescing neutron star and black hole binaries are among the most important targets of ground-based detectors, and a detector in space could obtain important information about a large variety of massive binaries.

Investigation 22.2. Gravitational waves from the Hulse-Taylor binary pulsar system

We will calculate here the wave amplitude that we expect from the Hulse-Taylor binary pulsar system PSR1913+16 and the energy it is radiating. The sizes of these numbers may surprise you!

For simplicity, we will consider here only binaries in which the masses of the stars are equal (call this M) and their orbits circular. Then because the two stars have the same mass, they also follow the same orbit, always lying opposite each other on a circle whose radius we will call R .

For a binary system with a circular orbit, the frequency of the gravitational waves is twice that of the orbit. The factor of two arises from the simple fact that, after half an orbit, the stars have replaced each other, and the gravitational field of the system is basically back to where it was at the beginning. So half an orbital period is a full gravitational wave period. This is true even if the stars do not have exactly the same mass, because the source of the gravitational waves is the elliptical asymmetry in the mass distribution, which is the same if one exchanges the two stars.

When we studied circular orbits in Investigation 3.1 on page 22, we found Equation 3.3 on page 22, that the acceleration of a body following a circular orbit of radius R with speed V is $a = V^2/R$. In the binary, this acceleration is produced by the gravity of the other star, which is a distance $2R$ away, so it is $a = GM/(2R)^2$. Setting these two expressions for a equal to one another tells us that the orbital speed is given by

$$V = \left(\frac{GM}{4R} \right)^{1/2}. \quad (22.5)$$

One gravitational wave period is the time it takes for one star to go halfway around the orbit, which is a distance of πR . At the speed V , this takes a time

$$P_{\text{gw}} = \frac{\pi R}{V} = \left(\frac{4\pi^2 R^3}{GM} \right)^{1/2}.$$

The gravitational wave frequency is the reciprocal of this:

$$f_{\text{gw}} = \frac{1}{2\pi} \left(\frac{GM}{R^3} \right)^{1/2}. \quad (22.6)$$

The amplitude of the radiated gravitational waves depends on the elliptical part of the kinetic energy of the system, projected onto the ellipses in Figure 22.1 on page 312. If we look down the axis of rotation of the orbit, then all the kinetic energy is in the plane of the sky. At one moment in the orbit the stars are moving in the x -direction in opposite senses, and a quarter of an orbital period later they are moving in the y -direction again in opposite senses. This is exactly what the test particles of the "+" pattern (top row) in Figure 22.1 on page 312 do, so all the kinetic energy of the stars contributes to the amplitude for this polarization. For two stars the total kinetic energy is $K = MV^2$, and we can use Equation 22.5 for V . Then we get for the amplitude along the rotation axis:

$$h_+^{\text{axis}} = 2 \frac{GM}{Rc^2} \frac{GM}{rc^2}. \quad (22.7)$$

The "x" polarization has the same amplitude up the rotation axis. It must, because the system is executing circular motion, and a simple rotation of 45° changes the "+" pattern into the "x" pattern:

$$h_x^{\text{axis}} = h_+^{\text{axis}}. \quad (22.8)$$

Exercise 22.2.1: Working out the algebra

Fill in the algebraic steps that lead to all the numbered equations in this investigation.

If we look at the system from a direction in the equatorial plane, then on average only half of the kinetic energy survives projection onto the plane of the sky, the rest being along the line-of-sight. And that half is only in the plane of the orbit: there is no circular symmetry from this viewing direction. So if we orient the plane along the x -axis in the viewer's coordinates on the plane of the sky, then the "+" amplitude will be half of its value on the axis, and the "x" amplitude will be zero:

$$h_+^{\text{plane}} = \frac{GM}{Rc^2} \frac{GM}{rc^2}, \quad h_x^{\text{plane}} = 0. \quad (22.9)$$

The amplitude expressions are based on a simple product of two terms. One of them, GM/Rc^2 , measures how relativistic the system is: how large the gravitational radius is compared to the orbital radius. The second, GM/rc^2 , is proportional to the Newtonian correction to the geometry of flat spacetime that produces the curvature of time and space for the Schwarzschild geometry.

The energy flux radiated by the system is given by Equation 22.4 on page 316, into which we can substitute the expressions above for the frequency and amplitude of the radiation to get (along the axis of rotation of the binary)

$$F_{\text{axis}} = \frac{\pi c^3}{4G} f_{\text{gw}}^2 (h_+^{\text{axis}})^2 + (h_x^{\text{axis}})^2 = \frac{c^5}{2\pi G} \left(\frac{GM}{Rc^2} \right)^5 r^{-2}.$$

In the equatorial plane this is reduced by a factor of eight.

What is of most interest normally is, how much energy is the system losing to gravitational waves? We can find its gravitational wave luminosity L_{gw} by adding up the flux radiated in all directions. If the flux were uniform in all directions, then it would be radiating the same energy per unit area per unit time across all parts of any sphere surrounding the binary. Taking the sphere to have radius r , we would just have to multiply the total area of this sphere, $4\pi r^2$, times the flux to get the luminosity.

The binary is not quite this simple, since the flux varies with direction. So we need to multiply the area of the sphere by the average flux. The flux in the equator is only one-eighth of the flux at the pole, but there is much more area near the equator than at the pole, so we might guess that the average flux should certainly be larger than one-eighth times the above expression, but possibly not as much as one-quarter times it. A full mathematical calculation shows that the correct factor is one-fifth. With this factor we get a formula that is actually the correct result for a binary whose orbit is basically governed by Newtonian gravity, despite the roughness of our derivation:

$$L_{\text{gw}} = \frac{2}{5} \left(\frac{GM}{Rc^2} \right)^5 L_E, \quad (22.10)$$

where $L_E = c^5/G$ is the Einstein luminosity, introduced in Chapter 21. It is striking how sensitive the binary's luminosity is to how relativistic the system is: the "relativity factor" GM/Rc^2 is raised to the 5th power, so a binary with just twice the orbital radius of another will radiate only 1/32nd (about 3%) of the energy. We expect to detect radiation only from the most compact systems.

more exotic are binaries in which two neutron stars are about to spiral together and form a single object. Two neutron stars will orbit one another hundreds of times a second in the last stages before coalescence, so the radiation will be observable by instruments built on Earth, if they are sensitive enough.

Such coalescing binaries are rare. The Hulse-Taylor system will spiral together about 100 million years from now. It is believed that there are a handful of other

Investigation 22.3. The shrinking orbit of the Hulse–Taylor binary pulsar system

Where does the energy radiated in gravitational waves come from? The stars themselves are not significantly affected: they retain their mass and size. The energy has to come from the orbital energy. We saw in Chapter 6 that the total energy of the orbit is conserved in Newtonian gravity. But gravitational radiation is not part of Newtonian gravity, and so the energy carried away by the waves results in a slow change in the orbital energy.

The orbital energy consists of two parts, kinetic and potential. They are actually closely related for the binary we are working with. We have seen that for the two stars together, $K = MV^2$. Using Equation 22.5, we get

$$K = GM^2/4R.$$

The potential energy of the two stars is the same as in Equation 6.9 on page 54, with m replaced by the mass of one of the stars and M_\odot by the mass of the other (both of which of course are M), and the radius r replaced by the distance between the two stars in the binary, $2R$:

$$V = -GM^2/2R.$$

The result is that the total binary energy is

$$E = -GM^2/4R. \quad (22.11)$$

This depends on the stars' masses and their orbital radius. The masses don't change as the binary emits gravitational radiation, so that the only thing that can change is R . Since the energy of the orbit must decrease by the amount that is radiated, it must become more negative, or in other words its absolute value must become larger. That means that R must become smaller. As R shrinks, the gravitational wave frequency f_{gw} given in Equation 22.6 increases. The signal is a whistle of gradually ascending pitch, which physicists call a **chirp**.

We can use these equations to deduce a characteristic time for the orbit to change. Let us ask how long it takes to cut the orbital radius R in half, doubling the absolute value of the energy. This means that the energy radiated must be equal to the absolute value of the energy at the beginning of this time. If the luminosity (the energy radiated per unit time) were constant in time, then the orbit-halving time t_{gw} would satisfy the equation

$$L_{\text{gw}} t_{\text{gw}} = GM^2/4R.$$

Of course, the luminosity is not constant, so this is not exact, but it should still indicate how long we have to wait for a substantial

change in the orbit. Using Equation 22.10 for the luminosity, we can solve for this characteristic time:

$$t_{\text{gw}} = \frac{GM^2}{4R} \frac{1}{L_{\text{gw}}} = \frac{5R}{8c} \left(\frac{GM}{Rc^2} \right)^{-3}. \quad (22.12)$$

This is called the **chirp time** of the binary. It is given by the light-crossing time of the orbit, $2R/c$, times a factor that is a sensitive function of how relativistic the system is. As the system shrinks, the chirp time gets shorter. This means that the chirp time is not very different from the full lifetime of the system: after the system has shrunk by a factor of two, it takes much less than the same amount of time again to shrink another factor of two, and so on until the stars coalesce.

In table Table 22.1 on page 319 we put some flesh on the abstract "bones" of all these formulas and evaluate the important numbers for three different equal-mass circular binary systems: a binary similar to the Hulse–Taylor binary today (but with a circular orbit), in our Galaxy at the distance astronomers calculate for the Hulse–Taylor system; a binary like the Hulse–Taylor binary at the time in the future when it is very near to coalescence, only placed at a distance of 200 Mpc from us, which is a distance where astronomers expect one such coalescence per year; and a binary consisting of two $10^6 M_\odot$ black holes at the center of a galaxy at a distance of 4 Gpc (which corresponds to a cosmological redshift of about $z = 1$). The implications of this table are discussed in the main text.

Now, when observing the gravitational waves, it is not usually possible to measure R or M directly: always measurable are h and f_{gw} , and if the system has a small enough orbital radius then t_{gw} may also be measurable. The properties of the system that determine these numbers are just the values of R , M , and the distance to the system r . These three unknown properties can be calculated if one can measure all three observables, since the three observables depend on R , M , and r in different ways. This leads to a profound result: one can measure the distance to a chirping binary just from the properties of the gravitational wave signal. **Chirping binaries are standard candles**. The distance can be estimated for any system whose orbit changes; it is not necessary to follow it all the way to the point where the stars coalesce.

We have demonstrated only that binaries consisting of equal-mass components in circular orbits are standard candles, but this important property actually extends to all binaries. Observed for long enough, the gravitational waves from any binary contain enough information to tell us how far away it is.

Exercise 22.3.1: Radiation from example binaries

Do the calculations that lead to the values in Table 22.1 on page 319 for the orbital numbers and chirp times from the values of M , R , and r given in the table.

Exercise 22.3.2: Chirp times

From the chirp time for the system that resembles the Hulse–Taylor pulsar that was calculated in Exercise 22.3.1, work out the rate of change of the period: what fraction of a second does the orbital period lose each second? Compare this with the measured number quoted in the text. Explain the difference. (See the caption for Table 22.1 on page 319.)

systems like the Hulse–Taylor system in our Galaxy that are too far away to be seen by today's radio telescopes, but if astronomers want to detect a few such events per year they must survey tens of millions of galaxies. This is a goal of present detector development.

What amplitude of radiation would we expect? From Table 22.1 on page 319 we find that we need to detect waves with amplitudes of a few times 10^{-23} . By comparison, the first laser-interferometer detectors, which are expected to begin operation in 2003, will have an initial sensitivity of around 10^{-21} .

Fortunately, physicists do not have to build detectors that are 100 times more sensitive than the generation now beginning to operate. The American physicist

Remember from Equation 22.1 on page 312 that the amplitude h of a gravitational wave is a dimensionless number.



Figure 22.5. Kip Thorne has had a major influence on the development of astronomers' understanding of black holes and gravitational waves, and he has been a driving force behind the development of interferometric detectors, which we will consider later. In particular he helped to found the LIGO project, which we will discuss later in this chapter. Drawing by Glen Edwards, used with permission.

Kip Thorne (b. 1940) was the first to understand how to benefit from the fact that binary radiation lasts for many cycles and is highly predictable. During the time the waves are in the observable frequency band, detectors will register tens of thousands of cycles of radiation. This will allow scientists to do **pattern matching** on the gravitational wave data, i.e. to look for a weak signal that matches the exact pattern of cycles that are predicted. Events will be reliably identified when they are only about ten times more sensitive than the first ones being built now (2002). This improvement in sensitivity is expected by about 2007, and frequent observations of coalescences of neutron stars can be expected soon after that.

Signals from binary black holes could be five to ten times stronger. But it is much harder to estimate the number of binary black hole coalescences that might occur. As we mentioned in Chapter 14, globular clusters may be efficient factories for binary black holes. It is possible that detectors might see many more coalescences of black holes than of neutron stars, or indeed that the first event detected will be a black hole coalescence. We will have more to say about this kind of observation below.

Merging black holes of larger mass are targets for the space-based detector LISA. We saw in Chapter 21 that the mean density of a black hole goes down as its mass goes up, so the orbital frequency near the horizon also goes down. Waves expected from holes between 1000 and $10^7 M_\odot$ are in the LISA frequency window.

The lower end of this range represents very interesting objects. Computer simulations have suggested that the first generation of stars to form, which were composed purely of hydrogen and helium, had much larger masses than we see in stars today, up to perhaps $1000 M_\odot$. Many or even most of these may have formed black holes, and surely left behind a population of binaries. LISA will be sensitive enough to see any systems in its frequency window anywhere in the Universe.

More massive black hole binaries may form from black holes that are in the centers of galaxies, as a result of galaxy mergers, as we noted in Chapter 14. LISA could again see any merger involving holes smaller than $10^7 M_\odot$ anywhere in the Universe. When one of the holes is much smaller, say $10 M_\odot$, LISA might be able to follow thousands of orbits before the smaller object crosses the horizon of the larger. These orbits would contain detailed information about the gravitational field outside the black hole, and from that information physicists could not only measure the mass and spin of the big black hole but even test the theorem that all black holes must have the Kerr geometry (Chapter 21).

Astronomers have another reason for searching for binary signals from systems whose orbits shrink during an observation. Such systems are said to “chirp”, because the frequency of their radiation increases with time. We show in Investigation 22.3 on the preceding page that such systems are *standard candles*: their waveforms encode their distance. If the radiation from chirping binary systems, including coalescing binaries, is observed by enough detectors to deduce the polarization and hence the intrinsic amplitude, then the systems will reveal their distance. This will be particularly interesting for, say, the $1000 M_\odot$ binaries that are seen at the time of the formation of the first stars.

Listening to black holes

In this section: gravitational waves form the only radiation emitted by black holes. Supercomputer simulations are needed to enable scientists to recognize the radiation.

Although astronomers are confident that many black holes can be identified using the techniques of optical, radio, and X-ray astronomy, all such identifications are indirect. They rely on electromagnetic radiation emitted by gas near the black hole. Apart from the impossibly weak Hawking radiation (Chapter 21), the only radiation that black holes themselves emit is gravitational radiation. When their horizons are distorted from their normally smooth shape by an interaction with another black

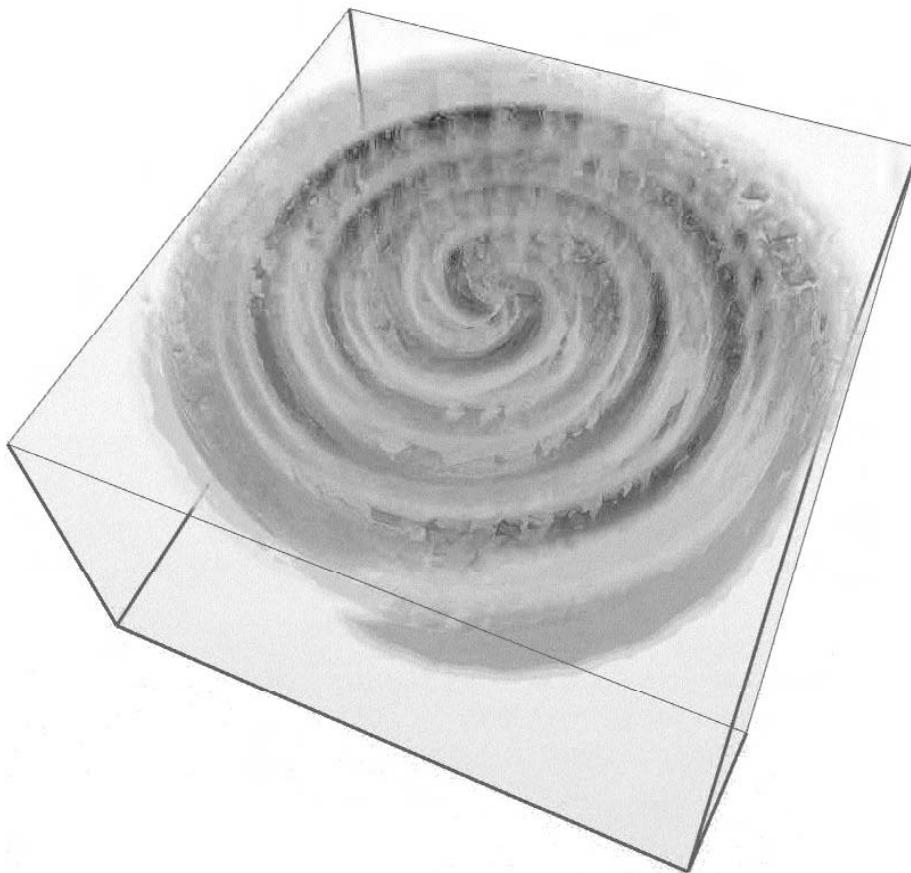


Figure 22.6. A snapshot of the gravitational waves emerging in the equatorial plane of a computer simulation of the merger of two black holes that have fallen together from a nearly circular orbit. The spiral nature of the waves reflects the orbital in-spiral of the two holes. The tightness of the spiral indicates how rapidly the black holes themselves are moving: since the waves move outwards at the speed of light, the holes must be moving at nearly that speed to wind the spiral so tightly. The waves carry away about 3% of the total mass-energy of the holes. Adapted from an image by W. Benger (ZIB), simulation by the Lazarus Project (AEI).

hole or a star, then the horizon wobbles for a short time, emitting gravitational waves until it settles into its quiescent state. Detecting this radiation, which has a recognizable signature, will be the first direct positive observation of a black hole. Astronomers will be *listening* to the holes themselves.

Detection of such events will not just be awe-inspiring. It will also test general relativity more stringently than any of its tests have done so far. The merger of two black holes to form a single one, with the emission of enormous amounts of gravitational radiation, is about as far from Newtonian gravity as one can get. But to perform this test, scientists have to make independent calculations of what radiation general relativity predicts such a merger to emit.

Such calculations cannot be done with pen and paper! In fact, they can't yet be done as accurately as will be needed, even with the fastest supercomputers available today (2002). But the next generation of computers may be big enough and fast enough to perform accurate simulations of the coalescence of two black holes. Teams of scientists around the world are working intensively towards this goal, and in many ways the work is as difficult and time-consuming as is the effort to build sensitive gravitational wave detectors. There are many teams of scientists collaborating on this problem around the world. Figure 22.6 illustrates a recent computation by one team of the gravitational radiation emitted in the equatorial plane by two in-spiralling black holes of equal mass.

Black hole collisions are pure Einsteinian gravity in action. No matter what kind

of a star collapsed to form the hole in the first place, the matter from that star is trapped inside (and probably turned into something like the state of matter at the beginning of the Universe!) and it cannot influence what happens to the outside of the hole any more. So when two holes merge, the result is independent of how they originally formed, and indeed it does not involve any matter at all. It is pure gravity in a vacuum. The merger is pure dynamical geometry.

It is eerie to think that thousands of stellar black hole mergers take place every year in the Universe, yet every event happens in complete silence apart from the whispers of the emitted gravitational radiation. These waves carry huge energies; a single stellar-mass black hole merger event has a gravitational wave luminosity greater than the luminosity in light of thousands of galaxies. Yet no-one has yet seen (or more properly, felt) a single event, and even the stars nearest the merging holes are hardly affected by the changes in gravity. Gravitational waves truly probe what Thorne has called the dark side of the Universe.

Gravitational collapse and pulsars

In this section: other potential sources of gravitational waves include gravitational collapse supernovae and pulsars with irregular mass distributions.

While binaries provide a huge variety of targets for gravitational wave detectors, there are other systems that could also be detected. The two that astronomers discuss most often are gravitational collapse and spinning, irregular neutron stars. Predictions about both are beset by many uncertainties.

Supernovae of Type II are rare and unpredictable events, occurring once in perhaps 50 years in any galaxy. Equally unpredictable is the radiation they will emit, because optical observations tell us little about how non-spherical the collapse and re-explosion will be. The best remedy for this uncertainty is to build detectors with great sensitivity. We have seen that an upper bound on the amplitude of this kind of radiation would be about 10^{-21} , and that is the sensitivity level of the first interferometric detectors, which will begin taking data in 2002. It seems likely, therefore, that first detections of supernovae will have to wait for the second-generation detectors.

However, gravitational wave astronomers must remain alert for such events. Gravitational waves are the form of radiation that will arrive first at the Earth from a supernova. If they can be recognized, then they will provide early warning to other astronomers that a supernova has occurred at a particular position, which should immediately be observed with other telescopes. The optical brightening of a supernova occurs several hours after the interior collapse, and has never been seen from the beginning.

The supermassive black holes in galactic centers may also have formed by gravitational collapse, whose radiation could be detected by LISA. This would, of course, help solve a number of mysteries about the origin of these ubiquitous objects.

Some spinning neutron stars could also be detectable sources. Unlike the narrow pencils of radio waves and light emitted by pulsars, any gravitational waves they emit would not be beamed. But the pulsar can nevertheless give off gravitational radiation if it is non-symmetrical about its axis of rotation.

Imagine a neutron star with a small lump on it somewhere. This could be a crack or deformation in the semi-solid crust of the star. Then as the star spins, the lump executes a circular motion not unlike the motion of the binary stars we examined above, and the radiation coming out will be similar. As for binaries, this radiation carries away energy. The effect would be to slow the pulsar down. Now, all pulsars are observed to be slowing down, but we have seen in Chapter 20 that this would be expected just from the electromagnetic radiation and relativistic particles they emit. Astronomers have no way of estimating how much of the slowdown to attribute

to gravitational waves. For any pulsar, the measured slowdown therefore gives an upper bound on the possible radiation from that pulsar. If the Crab pulsar radiates only 1% of its spindown energy loss into gravitational waves, it will be detected by the first generation of detectors within the first month or so.

Such a gravitational wave signal will be steady over long periods of time, so one's ability to find it increases with time, just as for the coalescing binary. However, the time-scale for achieving a detectable amplitude by pattern matching is many months rather than a few seconds. This makes extra demands on a computer-based analysis of the data. The reason for this is that the motion of the Earth, carrying the gravitational wave detector, introduces Doppler shifts into the observed frequency of the gravitational waves, and this exact pattern will have to be matched if the full sensitivity of the detector is to be realized.

One of the more exciting possibilities is the discovery of a previously unknown neutron star, just by the gravitational radiation it emits. For this reason, scientists want to make gravitational wave surveys of the entire sky. To do this, scientists will have to remove the Doppler shifts as for known pulsars. However, they won't know ahead of time the position of the neutron star, so they won't know what pattern to look for. A survey involves looking for all possible patterns. In a data set covering several months, this is such a complex job that it will require very fast supercomputers.

Neutron stars may also emit short-lived bursts of radiation from their normal modes of vibration, for example in the second or so after they are created, before they settle down into a quiescent state. This brief burst of radiation would be rich in information about the interior structure of the neutron star, in much the same way that the normal modes of the Sun have told us much about the solar interior. But these modes would be likely to radiate only weakly, so observations of this type are a long-term goal for more sensitive gravitational wave detector development.

Gravitational waves from the Big Bang: the Big Prize

To my mind the most exciting possibility of all for the new detectors is that they may be able to detect a background of gravitational radiation in space that originated in the Big Bang, the event that started the expansion of the part of the Universe that we can see.

We will learn, beginning in Chapter 24, that the Big Bang also produced a background of electromagnetic waves, with microwave frequencies: the cosmic microwave background radiation. By studying this radiation, scientists have been able to learn an immense amount about the early Universe, about the formation of the elements of which we are made, and about the formation of the galaxies that fill the observable Universe. Most important of all, these observations have given solid support to the idea of a Big Bang in the first place. Detecting gravitational waves from the Big Bang would offer fundamental information of a different kind.

The reason is that the gravitational waves would have been emitted much earlier than even the microwaves. Because gravity penetrates everywhere, the Universe is transparent to gravitational radiation, and has been so from the first moment. Electromagnetic radiation, on the other hand, was trapped and could not move freely in space in the earliest phase of the Big Bang, when matter was so hot that the whole Universe was an ionized plasma. Not until about 300 000 years after the Big Bang did radiation become free. So, the microwave radiation tells us about the Universe when it was a few times 10^5 years old. This is, of course, tiny compared to the present age of the Universe, about 10^{10} years, but it is still much later than the time at which some of the most interesting things happened, as we will see in the final chapters of this book.

>Of course, there is no lower bound: there could be no radiation at all!

In this section: the most fundamental observation that gravitational wave detectors could make is to measure random gravitational waves left over from the Big Bang. These would contain the imprint of the laws of physics at energies much higher than scientists can reach in Earth-bound accelerators. However, the task of detecting these waves is not simple.

Gravitational waves, on the other hand, were emitted just when all the interesting and poorly understood physics was happening, within a tiny fraction of a second after the Big Bang. To detect this radiation is to look at the Big Bang itself, and to get our first glimpse of physics at energies higher than physicists can ever hope to reach with Earth-based particle accelerators.

Unfortunately, scientists understand the physics of this very early period so poorly today that they have no strong predictions about the amount of radiation there is. Physicists have come up with very interesting mechanisms that could produce this radiation, and if it were found it would provide insight into physics at very high energies. But these models are speculative. What can be said with confidence is that the radiation today would just be a random background, looking like some extra noise in any one detector. Two detectors would be able to identify it, however, by looking to see if the noisy behavior of one is keeping step with the noisy behavior of the other. But physicists don't know yet whether they can expect this noise to be detectable by the instruments beginning their work now. The goal of detecting this radiation is likely to become more and more important in the design of detectors as they increase in sensitivity.

Catching the waves

In this section: gravitational wave detectors have been under development since the 1960s, and early claims of detections have been rejected. But the field today benefits from the ground-breaking work done by early researchers.

The last few sections have been a kind of menu of what the Universe might be offering us in gravitational radiation. Of course, any good menu also has the "chef's special", which is a dish you didn't expect when you went to the restaurant. Astronomers have to be alert for such things: in fact, most of the interesting discoveries of the last four decades in astronomy have been things that were not predicted or expected on the basis of prior knowledge. It would be strange if that did not also happen with gravitational waves.

So there is plenty of motivation for building detectors. But there is also plenty of reason to run away and do something else: trying to measure distortions in any man-made object at the level of one part in 10^{21} or even smaller is not a job to be undertaken lightly. In fact, it is a job that has taken many decades of work by a number of dedicated scientists, building detector prototypes that had only a small chance of detecting anything (such as rare supernovae in our Galaxy), gradually improving the technology until it was ready for the first generation of highly sensitive detectors, the ones that are under construction now.

The first gravitational wave detectors were developed by Joseph Weber (1919–2000) at the University of Maryland in the early 1960s. He considered a number of possible designs, and settled on what was the most practical for the technology of the time: a massive cylinder of metal, isolated as far as possible from external vibrations. When a gravitational wave hits this **bar detector** from its side, it induces a stretching and contraction along its length: just imagine the bar sitting along the horizontal diameter of the top row in Figure 22.1 on page 312. By instrumenting the bar to sense this stretching, Weber hoped to detect gravitational waves.

Weber's bar was one of the most sensitive instruments that had ever been built up to that time, but it was nowhere near the requirement of sensing a relative stretching of one part in 10^{21} . Probably his first bar did not do better than about one part in 10^{14} or so, although his subsequent instruments improved on this by perhaps a factor of 100. However, by the end of the 1960s Weber believed he was actually detecting gravitational waves, which were exciting two of his detectors simultaneously and frequently.

Many physicists responded to this extraordinary situation by building similar detectors. But detector after detector failed to confirm the observation. No other detector group found any significant excess of "events" over what they expected

▷This is called measuring a correlation between the noisy outputs of the two detectors.

from normal thermal and vibrational noise. Eventually the scientific community concluded that, whatever Weber had seen, they were not gravitational waves.

Many of the groups that built bars like Weber's decided to stay in the field, even though the rewards might be a long time coming. Some of them improved bar detectors by isolating them from external disturbances better and especially by cooling them to liquid helium temperatures (about 4 K), reducing their random vibrations and making weaker gravitational waves easier to detect. Today (2002) there are two bars, both in Italy, that are cooled below 0.1 K. These are called NAUTILUS and AURIGA. Since each bar weighs several tons, it is fair to say that these could be the largest such cold objects that have ever existed, even since the Big Bang! As their instrumentation improves, they should be able to detect broad-spectrum bursts near 1 kHz with an amplitude of perhaps 10^{-20} . Larger solid detectors, with a spherical shape, are now on the drawing boards. They could go a factor of ten or more better than this, breaking through the 10^{-21} barrier.

Such detectors will be expensive to make, and will not be ready until at least the second decade of the twenty-first century. The 10^{-21} barrier will be broken first, not by a bar, but by an interferometer. That is what we will look at next.

Michelson returns: the relativity instrument searches for waves

We have seen in Chapter 15 how interferometers work. An interferometer is designed to compare two lengths and to detect tiny changes in the difference between the lengths. To see how this can be used to detect gravitational waves, look again at our favorite figure in this chapter, Figure 22.1 on page 312. In the top row, place the center of the interferometer at the center of the circle, let the mirror at the end of one of the interferometer arms sit on the circle along the x -axis, and place the other arm's mirror on the circle at the y -axis. Now follow what happens to the lengths of the arms (the distances between the mirrors and the center) as the wave passes. Each arm changes length, but as one is expanding the other is shrinking. So the gravitational wave induces a change in the *difference* between the arms, and this is exactly what the interferometer is designed to sense.

The present interferometer projects grew largely out of decisions by some of the first bar-detector groups to explore interferometry as an alternative. These groups built interferometer prototypes and proved the technology would work. This led to the funding of much larger instruments, which could reach their initial goal of 10^{-21} by 2003. Large size is important for these detectors, to take advantage of the action of tidal forces, which produce length changes proportional to size. Since an interferometer measures changes in length by comparing them to the wavelength of the light it uses, a larger detector will produce a larger signal more easily than a smaller detector.

The most ambitious project is the Laser Interferometer Gravitational-wave Observatory (LIGO) in the USA, which is building two interferometers with arms 4 km long. One is at Hanford, Washington (see Figure 22.7), and the other at Livingston, Louisiana. At Hanford there will be an additional 2 km interferometer within the same system. LIGO could begin taking data of good sensitivity in 2003. On a similar timetable, but smaller in size, is GEO600, a detector being built near Hannover, Germany, by a German-

▷Weber himself continued to maintain until he died that something interesting had been exciting the detector, but he convinced few other scientists of this. The history of physics has other episodes of a similar nature, where results that are accepted for a time are later discarded, often unexplained. What counts to the majority of physicists is not whether each experiment can be explained: only if its results can be duplicated by other scientists does it demand to be accepted.

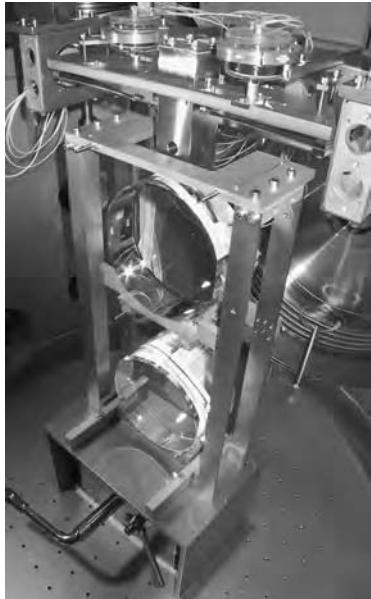
In this section: the first detections may be made by interferometers, descendants of Michelson's original instrument. We describe the principle of using these as detectors, the major projects that are building them – LIGO, VIRGO GEO600, TAMA300– and the kinds of challenges involved.



Figure 22.7. The LIGO detector at Hanford, in the US state of Washington. Each arm stretches 4 km, with mid-stations at 2 km. The arms house the world's largest vacuum system, inside of which intense laser beams monitor the separations between mirrors that are suspended in such a way that they are free to move along the direction of the arm in response to a gravitational wave. The system is able to sense motions as small as 4×10^{-18} m if they occur in 10 ms. Image courtesy of LIGO.

British collaboration of scientists. GEO600 is developing more advanced optical technology to achieve a sensitivity similar to that of LIGO with arms that are only 600 m long (see Figure 22.8). This technology, once proved in GEO600, will assist LIGO and the other larger detectors to upgrade their sensitivity later in the decade.

Figure 22.8. One of the mirror assemblies for the GEO600 detector. The bottom piece is the mirror, the upper disk a balancing part of the suspension. The mirrors are made of high-quality fused silica (quartz) and are suspended from fibers made of the same material. The quality of the suspension is such that if the mirror pendulum were set swinging and left alone in the vacuum system, it would swing for many months before the amplitude reduced by half. Image courtesy of GEO600.



can be defined to a high accuracy even though individual atoms move around by much more. However, to sense even an average displacement of the surface of a macroscopic object to this accuracy requires excellent mirrors and high-power continuous wave lasers. It also requires that the light move in a very good vacuum. The LIGO detectors have a vacuum whose volume is larger than any constructed before, even for the big particle physics accelerators. Finally, projects like these require money. LIGO is the largest single scientific project that the National Science Foundation in the USA has ever undertaken.

Developers of these instruments have had to learn how to cope with the same basic sources of instrumental noise as the bar detectors fight against: external vibration and internal thermal motions. These sources of noise set a lower limit on the frequency window at which they can observe. In this window, the sensitivity of an interferometer is limited mainly by how much light is used in the interferometer. This is because the mirror displacements caused by the gravitational wave are much smaller than a wavelength of light (most detectors use infrared light with a wavelength of about 1 μm), and the precision with which they can use light to pinpoint small displacements depends on how many photons are used. Each photon arrives, by the uncertainty principle, with a randomness that makes the interference slightly “fuzzy”. The more photons one uses, the less the average randomness is. This fuzziness, called **shot noise**, is the third limiting factor on the sensitivity of interferometers.

The fact that several interferometers are under construction is not redundancy. As we noted above, the data from three or four interferometers observing a given source simultaneously are necessary to locate the position of the source in the sky and determine the polarization of the incoming waves.

Only slightly smaller than LIGO is the single 3 km Italian–French detector VIRGO, being built near Pisa. VIRGO should begin operations about one year after LIGO and GEO600. In Japan the TAMA300 detector in Tokyo, with 300 m arms, is a prototype for a later Japanese detector with 3 km arms, which is planned to leapfrog the initial LIGO and VIRGO detectors and go straight to second-generation sensitivity. There is a proposal in Australia for a detector called AIGO that would have even longer arms than LIGO, but as of 2002 this had not been funded.

Even over 4 km, a disturbance of 10^{-21} translates into a mirror motion of 4×10^{-18} m, less than the size of a proton. Of course, such instruments do not measure the positions of single protons; rather they sense the average position of the surface of a mirror, which

This is a field that is developing rapidly. By the time you read this, the major instruments may already be operating. You should keep your eyes on the scientific press for further developments.

LISA: catching gravitational waves in space

We have seen that the frequency range around 1 mHz is very interesting for gravitational-wave astronomy. But it is not possible to observe this low-frequency band from the Earth, no matter how much scientists improve the technology. The reason is that the Earth itself creates gravitational noise that is stronger than the signals astronomers expect from these sources. A gravitational wave detector is simply an instrument that responds to tidal gravitational forces, of whatever origin. When a heavy truck drives past a laboratory, its small gravitational field can be much larger than a weak signal from space. If the vehicle takes 30 s to drive past, then its effect will confuse wave observations in a frequency band of $1/(30\text{ s}) = 33\text{ mHz}$.

This has nothing to do with mechanical vibrations: it is gravity itself, and it can't be screened (Chapter 1). The only remedy is to get far from the Earth, because these disturbances get weak very rapidly as one goes further away, while the size of gravitational waves from very distant astronomical sources essentially doesn't change.

Space-based searches for gravitational waves have already been made using communication signals between the Earth and interplanetary space probes. To track spacecraft in the Solar System, space engineers continually send out radio signals to them and receive signals back. If the radio waves were simply reflected from the spacecraft, as from a mirror, they would be too weak to detect when they returned to the Earth. Instead, space probes carry **transponders**, which are systems that receive a radio signal from the Earth and re-transmit back to Earth an identical signal; they effectively act as amplifying mirrors. Now, a passing gravitational wave would affect the time it takes radio waves to travel out to the probes and back; the signature of a gravitational wave in such data is unique, and sensitive searches can be made at very low frequencies. No positive detections have been reported so far, and it is unlikely that sensitivity will be improved in the near future to levels below $h = 10^{-16}$. This is not good enough, as we can see from Table 22.1 on page 319.

The LISA detector is designed to do better than this by roughly a factor of 10^6 . Developed first for ESA by a team of European and American physicists, the mission is now a joint project of ESA and NASA, scheduled for launch in 2011. LISA will consist of three spacecraft orbiting the Sun in the Earth's orbit, about 20° behind the Earth. They will form a roughly equilateral triangle with arm lengths of about $5 \times 10^6\text{ km}$, many times larger than the the Sun. Of course the arms would be empty: space is already a good vacuum. But laser light would be transmitted from one spacecraft to another and back, between all three pairs. The three would form two essentially independent interferometers, which would together extract all the desired information from the waves. The arrangement chosen for the spacecraft insures that, as they follow their individual elliptical orbits around the Sun, their pattern remains an equilateral triangle that rotates once per orbit (see Figure 22.9). The mission could make observations for a

In this section: the most spectacular gravitational wave detector being planned is LISA, a space mission to detect low-frequency gravitational waves. Three spacecraft will orbit the Sun and measure tiny changes in their separations. The launch is planned for around 2011.



Figure 22.9. An artist's view of how LISA would look in orbit around the Sun, about 20° behind the Earth. The three spacecraft follow free orbits around the Sun, chosen so that they remain in formation, always facing the Sun and lying in a plane tilted at 60° to their orbits. The sizes of the spacecraft, Sun, and Earth are, of course, not drawn to scale. Image by Chris Osland and Jonathon Copeland, Rutherford Appleton Laboratory, used with permission.

decade or more.

The technology of LISA is fascinating in itself. The spacecraft will contain small cubes (called proof masses) that are the reference points for the interferometry. These must remain undisturbed so that they respond only to gravitational waves. Accordingly, they will fly freely inside the spacecraft, which will continually sense their position and fire tiny retro-rockets to counteract any external forces on the spacecraft (such as from variations in the radiation pressure of sunlight). The job of the spacecraft is to shield the proof masses so that it does not bump into them!

LISA also does its interferometry in a different way. Its “mirrors” are not reflectors, as in ground-based interferometers. Instead, they are transponders, just like the systems carried by interplanetary space probes to allow them to be tracked from the Earth. On LISA, however, they transpond laser light rather than radio waves.

LISA will make observations that are complementary to those made by ground-based detectors. In fact, LISA has been designed mostly by scientists who also work in ground-based projects. It will have sufficient sensitivity to see waves from black hole collisions wherever they occur in the Universe. It will also be able to do a census of binaries in the galaxy in its frequency range. Moreover, it will also look for a cosmological background of gravitational radiation in its frequency range, with a sensitivity comparable to that of the ground-based detectors, which operate at a much higher frequency. Even if it does not see the background, scientists and the agencies are already studying LISA follow-on missions dedicated to detecting the Big Bang. Gravitational wave astronomy is a field that can be expected to have a long future in space.

Gravitational lenses: bringing the Universe into focus

As we have progressed through the story set out in this book, we have met and begun to understand many of the objects that astronomers regularly photograph: planets, stars, galaxies, supernovae. Astronomical photographs show, in fact, the astonishing variety of objects that make up our Universe. But, to my eye, the most spectacular and entertaining astronomical photographs are fashioned by the objects we will study in this chapter: gravitational lenses. Let's start this chapter with two, shown in Figure 23.1 on the following page. Gravitational lenses are a spectacular illustration of the working of general relativity in the Universe. And besides entertaining us with pictures of eerie beauty, they have become an important tool of astronomy, a way of probing the distribution of mass (and in particular the dark matter) in galaxies and clusters of galaxies.

Lensing can also be a nuisance, of course. Imagine trying to count the number of galaxies that existed at some early time in the evolution of the Universe, to try to pin down the details of galaxy formation, and being confronted with the second photograph in Figure 23.1. How do you know how many images of distant galaxies correspond to only a single galaxy? How do you know if the images are brightened by the lens, so that if the lens had not been there you would not have seen the galaxy in your survey at all? (In fact, the photograph in Figure 23.1 came from just such a project!) Whether lensing is a nuisance or a tool, whether an astronomer wants to remove its effects or use them to study other things, gravitational lensing is important, and it is one of the big research areas in astronomy today.

Pretty obvious, really, ...

Gravitational lensing is a direct consequence of the fact that light is deflected by gravity. The Cavendish–Soldner–Einstein derivation of light deflection, using only the equivalence principle (see the discussion in Chapter 4), is itself enough to establish the principle, although to do quantitative studies it helps to know (Chapter 18) that the deflection predicted by general relativity is twice that due just to the equivalence principle alone. That was established by the eclipse expedition of Eddington and Dyson in 1919.

Nevertheless, gravitational lensing did not become a serious part of astronomy until 1979, when two images of a distant quasar were first identified using radio observations at the Jodrell Bank radio observatory in England. During the intervening years, much of the basic theory had been worked out, but the chances of observing lensing were thought by most astronomers to be small. This was partly because astronomers had at that time no clear idea of how much mass there is in the Universe, and they had no idea how clumpy it would turn out to be. Clumps make better lenses than smooth distributions.

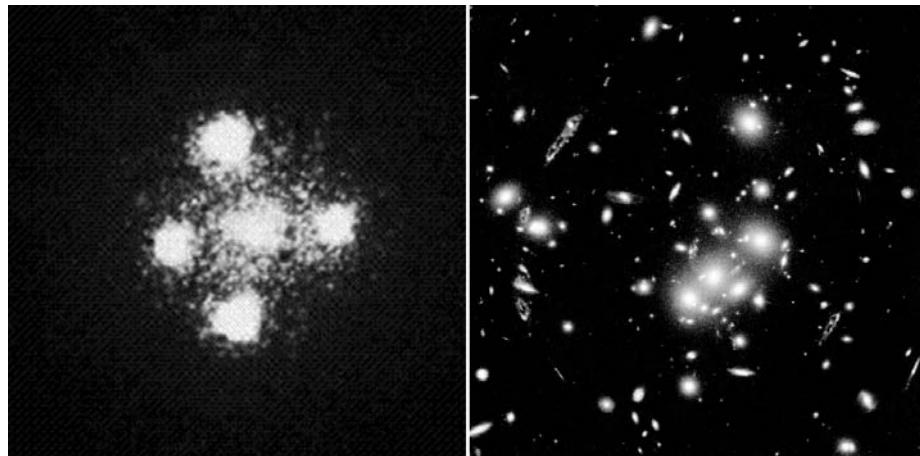
The lack of interest in lensing also reflected the technology of the day. In many lenses the image separation is less than one arcsecond, which is difficult for ground-based optical telescopes to resolve. Moreover, to see lensing of this angular size,

In this chapter: gravitational lensing has become one of the most important tools astronomers have for investigating the true distribution of mass in the Universe, and for measuring the Hubble expansion rate. We study how lensed images form, why lenses produce multiple images (always an odd number), why some are magnified, and how lensing and microlensing are used by astronomers.

>The image under the text on this page is of the gravitational lens known as B1359+154, which is remarkable for showing six different images of the same distant galaxy. The circled images are the three galaxies that create the lens, at a cosmological redshift of about one. The remaining images represent the same very distant galaxy, which has a redshift of 3.2. (HST image courtesy STSCI/NASA.)

In this section: lensing could have been predicted hundreds of years ago. When it was finally observed it came as a surprise.

Figure 23.1. Two photographs of gravitationally lensed images. The first image is known as the Einstein Cross, or g2237+305: four perfectly symmetrically placed images of a single distant quasar around the central core of a spiral galaxy. The second is a complex and rich cluster of galaxies (the fuzzy objects) called CL0024+1654, that form a lens that produces at least five different images of a single much more distant spiral galaxy. (Left image courtesy Goddard Space Flight Center (GSFC), HST, and NASA. Right image taken by W Colley, E Turner, and T Tyson, HST/STSCI/NASA.)



the lensed object should have a very small angular size as well, or all the effects will get washed out in its large image. Until the discovery of quasars in the 1960s, astronomers thought that the only objects distant enough to be lensed by galaxies would be other galaxies, and their large angular size would make lensing hard to identify. But quasars are different (as we saw in Chapter 14): bright enough to be seen even at great distances, they are nevertheless small enough that their appearance is point-like.

It is not surprising, therefore, that the first lensed object to be discovered was a quasar, and that it was discovered with a radio interferometer, which is an array of several radio telescopes that, when observing the same source together, have excellent angular resolution. The two images of the quasar were separated by only 6 seconds of arc. It is also not surprising that the discovery was completely by chance: the observers were just doing a survey of quasars. When they found two quasars unexpectedly close together, they went to an optical telescope and took spectra of both objects. Quasar spectra are like fingerprints: although they all share some general features, no two are exactly the same. In this case, however, the two spectra were identical. Astronomers had their first gravitational lens.

Since that time, hundreds of gravitational lenses have been discovered and studied. Lensing by huge clusters of galaxies and by individual stars smaller than the Sun has been detected. Gravitational lensing is teaching scientists about the dark matter on all length-scales, from within our Galaxy to within clusters of galaxies.

...but not always easy

In this section: the observation and interpretation of gravitational lenses present real challenges.

Although the basic ingredient of an analysis of lensing is just the equation for the relativistic deflection of light in a Newtonian gravitational field, Equation 18.13 on page 235, the use of this formula to understand realistic lenses requires elaborate calculations on computers. The main problem is that Nature provides us with very complicated lenses to interpret. A telescope maker on Earth spends enormous effort to make the mirrors and lenses of a modern telescope smooth and perfectly shaped to small fractions of the wavelength of light, and then the telescope is used to observe light that has come to us through a bumpy, astigmatic, partly absorbing natural gravitational lens!

The principles of lensing are not difficult, however, and in this chapter we will concentrate on understanding simple lenses. We will discover the peculiar nature of the gravitational lens, divergent in some regions and convergent in others; we will see why lenses magnify objects and make them brighter; we will see why there are

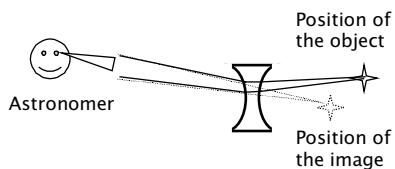


Figure 23.2. A diverging (concave) lens is placed between an astronomical source and the observer (with telescope). Because the rays are spread apart as well as being deflected, when the astronomer receives them they seem to be coming from a nearer source, in a different direction. The location of this image is drawn with lighter lines. It is the direction the astronomer has to point the telescope in order to see the source.

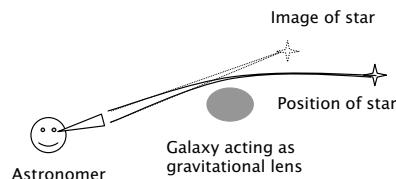


Figure 23.3. A galaxy deflects light as it passes it on its way from a star to an astronomer. The angular position of the image is further from the lens than the true position of the star. Moreover, because the light ray nearer the galaxy has bent more than the other one, they are diverging faster when they reach the telescope than they would have been if the galaxy had not been in the way. This makes the image appear closer.

in principle always an odd number of images of any lensed object, although not all of them will necessarily be bright enough to be detected; and we will see how lensing can be used to measure the mass of the lens itself and, possibly, the expansion rate of the Universe.

How a gravitational lens works

The most spectacular photographs of gravitational lenses are of systems with multiple images, as in Figure 23.1, but lensing also operates on single images, and it is easiest to start with that. We will see that a gravitational lens works, most of the time, as a **diverging lens**. To see what that means, we first look at the way a diverging lens made of glass works. In Figure 23.2, we imagine a glass lens between a star and an astronomer. The lens is concave. When a parallel beam of light passes through such a lens, it spreads out. This is why we call it a diverging lens. The spreading of the beam has a very interesting consequence: it leads to magnification. Consider a pencil of light rays emanating from one point on the star in the figure, and passing through the lens before arriving at the telescope. The pencil is diverging anyway, and in principle it would be possible to measure the distance to the star from the angle the diverging rays make. You just trace them back to where they intersect, and that must be where they came from. This is how the parallax method of determining astronomical distances works, as we saw in Chapter 4. This is also how our brains judge distances, using this divergence information as recorded by our binocular eyes.

When the rays pass through the glass lens, they get a bit of extra divergence, as well as an overall change of direction. When they arrive at the telescope, the astronomer infers where they came from by tracing straight lines back along the incoming rays. How this works is shown in the figure. The rays intersect in the “wrong place”, of course. This is the *image location*, the place where the lens has fooled us into thinking the object is. Because the lens has made the divergence stronger, the image location is closer than the real location. The object appears closer than it really is. It follows that it looks larger and brighter than it would without the lens: it is magnified.

Now, the same thing happens with a gravitational lens, but with a rather different geometry. In Figure 23.3 we see a single galaxy acting as a lens. Instead of pushing the light away from the axis (the line between the astronomer and the center of the lens), as in the case of the glass lens, gravity pulls the light towards the axis. But the pull of gravity is stronger on the rays that are closer to the axis, closer

In this section: how gravitational lenses form images. How they work as diverging lenses, and why the divergence makes the image brighter.

to the galaxy. So these get bent more, and the result is that the whole beam is given a little extra divergence.

The net effect is the same as with a glass lens: the extra divergence makes the image appear closer, and therefore brighter. The figure shows that the overall bending of the path of the light moves the image *away* from the galaxy that acts as the lens.

Why images get brighter

In this section: we explain why a diverging lens makes images brighter. This applies to glass lenses as well as gravitational lenses.

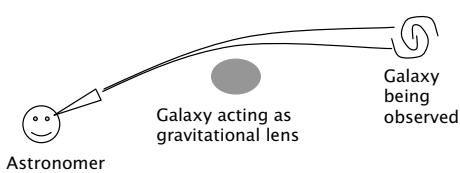


Figure 23.4. The brightness of the image depends on how much light from the source arrives at a point in the astronomer's telescope. Therefore we draw light rays that originate at the astronomer, the opposite of the rays we drew for working out where the lensed image was, in Figure 23.3 on the previous page. The divergence of the rays means that they cover more of the surface of the object (a galaxy in this diagram) than they would have covered if the lens had not been there. This compensates the divergence of the rays from the galaxy, so that the brightness of any small angular part it is the same as if the lens were absent.

This apparent paradox is present in the theory of the glass lens, too. Its resolution is to realize that the brightness of the image is not represented by the rays in Figure 23.3. They show the light from one point on the star as it reaches many places in the telescope. The brightness of the image, on the other hand, is determined by the rays that reach a given place on the telescope from different places on the star: how much light do they bring from the star?

As a first step in understanding what happens, we discuss what happens when the observer looks through the telescope at a distant galaxy, rather than a tiny star. Consider a small pencil of rays from the observer that reach the galaxy, as in Figure 23.4. Suppose in fact that the pencil is so narrow that when it reaches the galaxy it covers only a small part of the surface of the galaxy. (These are the rays that, say, will bring the light to one **pixel** of the image of the galaxy on the observer's photographic plate.) Since the lens has made the pencil diverge more than it would have if the lens were not there, these rays intersect more of the surface of the galaxy than if the lens were not there. This tends to bring *more* light into the observer's eye. In fact, it exactly compensates the divergence we noted in the first paragraph of this section: the light from the surface of the galaxy is indeed being spread out more by the lens, so less of it reaches us from any part of the galaxy. But the lens allows us to fit more of the surface of the star into our pencil of rays, with the following net result.

A pencil of rays with a given angular width receives the *same* amount of light from the galaxy regardless of whether the lens is present or absent, provided that the pencil is smaller than the angular size of the galaxy.

Naturally, this is true only if the lens is transparent; we don't worry here about absorption or scattering of the light by the lensing objects.

Therefore in Figure 23.4 we draw the same situation, but we trace rays back from the astronomer to the star. They pass the galaxy and are lensed in exactly the same way, which means they are given a little extra divergence. The effect of this is that when they reach the star, they occupy *more area* on the star than they would have if the galaxy had not been there.

The extra brightness of the image of the star comes from the fact that more of the star is contributing light to this point at the entrance to the telescope: the image of the star is brighter because more light from it arrives at the telescope than if the lens were not there.

The word that astronomers use for the amount of light received from a piece of the surface of an object into a given angle at the telescope (into a given pixel)

is **surface brightness**. We have shown that the surface brightness of an object is unchanged by the lens. This applies to lenses in ordinary optics as well as to gravitational lensing.

But still, why do stars look brighter though a gravitational lens? Have we not just proved that they will be the same brightness? No, we have proved that a piece of the star, covering a given angular size in our observation, will be just as bright as before. But when we consider the whole star, we need to take account of the fact that the size of the image of the star is larger than without the lens, because the diverging lens has made the star appear closer. The star occupies a larger angular size on the sky. Since it has the same surface brightness, we get more light in total from it.

This effect is particularly important for stars, whose angular sizes are so small that astronomers cannot resolve individual parts of their surfaces. In a photograph, it is not possible to tell that the star is in fact larger, since its size is still too small for the telescope to resolve. All we see in the photograph is that there is more light from the star: it is brighter. Galaxies that are big enough to resolve in a photograph, on the other hand, are no brighter in a given area of the photograph than without the lens. They are simply bigger.

You might now ask, what happens to conservation of energy? If there is more light arriving at the telescope from the star, where is this energy coming from? Clearly, the star is not making any extra light, nor is the lens. The extra light in the telescope is light that would have gone elsewhere but is being re-directed by the galaxy into the telescope. Therefore, some other astronomer must be losing the light that should have arrived. Where is he?

The position of the astronomer in Figure 23.4 is not particularly special. Any astronomer will get a little extra light if the light passes the galaxy. The astronomers who lose out are on the other side of the star. The galaxy's gravitational attraction has pulled a little of the light that is meant to go to the right in the diagram and is sending it to the left, and this allows the re-distribution of light that makes the image brighter.

If the lens were more complicated, like the one we are about to look at in Figure 23.6 on the next page, then the situation is also more complicated. Different parts of an image can brighten up at the expense of neighboring parts, as well as at the expense of the unfortunate astronomers in the other half of the Universe.

Making multiple images: getting caustic about light

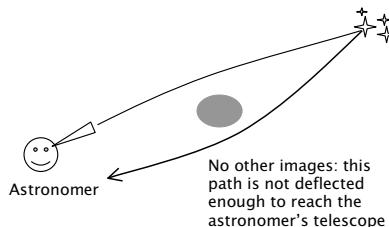
If you bought a new camera and found, when you had developed the first roll of film, that there were two images in the same photo of your grandmother holding her pet dog, you would feel cheated and you would demand your money back. But when gravitational lenses do this, we all get excited! In fact, we will see that gravitational lenses typically give you *three* images of your grandmother, and in one of them she is left-handed!

In this section we explore one of the extra images, called the second direct image. Its existence is easy to understand and of most interest for applications of lensing. In the next section we will investigate the Einstein ring, an important radius around the lens, which is the key to understand microlensing and the discovery of MACHOs (see page 172). After that we will thread our way through the subtle reasoning that shows that if there is a second image then there is also a third, in which left and right are reversed.

What normally happens is that there is only one image, a little brighter. Other rays, that pass on the other side of the lensing galaxy, do not deflect enough to reach the astronomer. This is illustrated in Figure 23.5 on the following page. Notice,

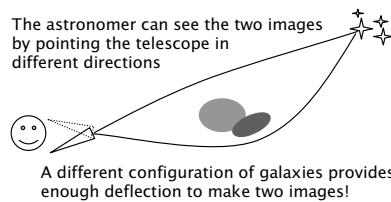
In this section: we learn that the number of images is related to the way the light rays from the object intersect one another in caustics.

>Of course, if your grandmother really is left-handed, then in the third image she will be right-handed!



No other images: this path is not deflected enough to reach the astronomer's telescope

Figure 23.5. Every galaxy (or any other mass) produces a small deflection of light rays, but for multiple images to form there has to be another path that brings light from the star to the astronomer. In this case, light going around the other side of the galaxy is not deflected enough to reach the astronomer.



The astronomer can see the two images by pointing the telescope in different directions

A different configuration of galaxies provides enough deflection to make two images!

Figure 23.6. If another galaxy is put into the lens, gravity may now be strong enough to deflect light on a second path from the star to the astronomer. This light arrives from a different direction, so the image looks like a distinct object. But it is just a different view of the same star.

however, that the rays that pass on either side are approaching each other. They just don't have time to meet before one of them meets the astronomer. If the astronomer goes much further away, then she can find a location where both rays arrive at the same location, and she will see two images. It is important to realize that for gravitational lenses, as for any other kind of optics, the location of the observer is as important for determining what is seen as are the locations of the source and the lens.

Suppose the observer is very close to the lens, so close that there is only one image of a distant star, as in Figure 23.5. The other way to get two images is to modify the lens. In Figure 23.6 we have added another galaxy with a lot more mass to the lens. This bends the light from the star much more on that side of the lens, and directs it toward the astronomer. The astronomer can now see the two images by pointing her telescope in the two different directions.

For the astronomer to see multiple images, she must be sitting in a special region, where light rays emitted from the source in different directions intersect one another. The nice smooth light-cones of special relativity have no self-intersections. But even a small amount of matter in spacetime will distort light-cones enough to make them fold over on themselves. The boundaries of regions in which self-intersections occur are called **caustics**. To see multiple images, there needs to be a caustic between the observer and the lens.

Then let the observer move further away. When she encounters the first caustic, she will see three images instead of one. In Figure 23.6 it is clear that there are at least two images. We will see why the number actually goes up to three later in this chapter.

If the lens has a complicated shape, she might encounter another caustic if she moves even further away, or to the right or left. At this caustic, two more images appear, giving a total of five. Even more are possible.

The Einstein ring

In this section: we learn about the characteristic size of the lens, called the Einstein radius. For symmetrical lenses, an object directly behind it is lensed into a ring, called the Einstein ring. In more general lenses, all secondary images appear within this radius.

There is an important length-scale in gravitational lensing, called The **Einstein radius**. It is defined by assuming, just for the purpose of the definition, that all the mass of the lensing object is arranged in a perfectly spherical way, and the object being lensed is point-like, is directly behind the lens from the astronomer's point of view, and is at just the right distance for its light to bend enough to reach the observer. This is illustrated in Figure 23.7 on page 338. With this simplified geometry, the light from the point-like "star" is not focused into a single image, but is spread

►Figure 23.6 is highly exaggerated, of course. Normally astronomers see images with only small separations, and both images are typically in the same field of view of the telescope. But it is better to draw these exaggerated views when trying to understand the phenomenon, to avoid confusion between the rays.

Investigation 23.1. The Einstein radius and ring

Our aim here is to use the geometry of the triangle drawn in Figure 23.8 on the following page to calculate the distance of closest approach b to the lens. The trajectories of the light rays are approximated by straight lines in this calculation, but this is not a bad approximation for rays that have to travel vast distances between star, lens, and astronomer.

The key to the calculation is that light is deflected by the Einstein formula, and this gives the angle θ between the direction the light was going before it encountered the lens and the direction it is going after leaving the lens, given in Equation 18.13 on page 235:

$$\theta = 4GM/c^2 b.$$

This angle is, by the geometry of triangles, the same as the sum of the angles $\alpha + \beta$. Each of these is simply expressed in terms of the distances shown. Since in any reasonable case, the angles will be very small and so b will be very small compared to the other distances, we can use the small-angle approximation to the tangent

function and write

$$\alpha = b/D_L, \quad \beta = b/D_{LS}.$$

Then the equation $\theta = \alpha + \beta$ is

$$\frac{4GM}{c^2 b} = \frac{b}{D_L} + \frac{b}{D_{LS}}.$$

Solving this for b gives the result quoted in the body of the chapter, Equation 23.1. Dividing it again by D_L gives the quoted result for the angle α that the astronomer measures for the ring.

Actually, our calculation is valid only if the lens and source are nearby. If the distances D_L and D_{LS} are large on a cosmological scale (see the next chapter), then the curvature of spacetime means that they can't simply be added together in Equations 23.1 and 23.2 to give the overall distance from the observer to the source of light (the star). Instead astronomers use the symbol D_S for this.

Exercise 23.1.1: Einstein ring

Perform the indicated algebra to derive the Einstein radius and its angular size.

out into a ring, coming to the astronomer from all directions around the lens. This ring is called the **Einstein ring**. Even though real lenses are not usually so symmetrical, the size of the ring is an important measure for the lens. It is roughly the place where one can expect strong magnification of images.

Nature comes surprisingly close to producing an Einstein ring image once in a while. The Einstein Cross in Figure 23.1 on page 332 is an image produced by a galaxy with a roughly elliptical shape exactly in front of the true position of the quasar. The elliptical shape of the galaxy pushes the light into four images, rather than allowing it to spread evenly over the ring. The radius of the circle on which the images lie is the Einstein radius. In Investigation 23.1 we show, using simple trigonometry, that its radius is, for a lens of mass M ,

$$b = \left(\frac{4GM}{c^2} \frac{D_L D_{LS}}{D_L + D_{LS}} \right)^{1/2}. \quad (23.1)$$

All the symbols in this equation are defined in Figure 23.8 on the next page. From the astronomer's point of view, what is important is the angular size of the ring, which (when measured in radians) is this distance b divided by D_L :

$$\alpha = \frac{b}{D_L} = \left(\frac{4GM}{c^2} \frac{D_{LS}}{D_L(D_L + D_{LS})} \right)^{1/2}. \quad (23.2)$$

In this configuration, a great deal of the light from the star is directed toward the astronomer. It is as if there were images of the star all around the ring instead of just in one place. If the astronomer is able to resolve the ring, as in the Einstein Cross, then she will see a ring or, more realistically, a series of images or beads arranged around the ring. Each will be as bright as a single lensed image of the star. If the ring is too small to resolve, then the astronomer will just see a very bright single image.

Now, perfect alignments of the lens with the source and astronomer are rare. So let us ask what might happen if the distant star or quasar is not inside the Einstein radius around the galaxy. Will there be any lensing at all?

Yes, because it is clear that light can approach closer to the lens than the Einstein radius. So if a star is displaced to the side, then to form a second image the light has

to bend more on going past the galaxy. It can in principle do this by approaching closer to the galaxy. So stars outside the Einstein radius can still form secondary images if the lens is compact enough. However, because the light ray comes in closer to the galaxy, the apparent position of the *image* will be inside the Einstein radius. Thus, we have the following important result.

All the secondary images produced by a spherical gravitational lens will be within the Einstein radius, even if the position of the object itself is outside this radius.

Now, of course real lenses are not perfectly spherical, and so our definition of the Einstein radius cannot be exact for more realistic lenses. But it serves as a good guide to the region in which one should look for images. And if you see a number of images of an object, then it is possible to estimate the Einstein radius from them. If these images are distributed around a circle, then you can be confident that the circle is the Einstein ring and the object itself is inside that ring too.

MACHOs grab the light

In this section: microlensing studies, where one looks for lensing by small objects, indicate that the halo of the Galaxy is composed partly of dark, compact objects.

One of the most interesting recent applications of lensing is to the search for the dark matter in the Galaxy. It is possible that at least some of this matter is in large objects, of the mass of Jupiter or larger, but which are not massive enough to initiate nuclear reactions and give off light. These are called brown dwarfs: *brown* because they don't radiate much light, and *dwarfs* because they are small stars. It is also possible that there are other small, dark objects of an unexpected nature, such as boson stars (Chapter 12). The generic, and somewhat whimsical, name that astronomers use for all of these is MACHOS, which we first met in Chapter 14.

Although dark, MACHOS could be detected if they act as a gravitational lens on a more distant star. The idea is that the MACHO would be moving in the halo of our Galaxy, and would pass in front of a distant star by random. The star would brighten up temporarily, and then as the MACHO moved on it would return to its original brightness. To detect MACHOS this way requires extensive monitoring of millions of background stars, waiting for chance events. It also requires painstaking investigation of each candidate event, to insure that it was not caused by something else, such as a variable star.

Let us first look at some typical numbers. If a MACHO is in the halo of our Galaxy, it might be a typical distance of 5 kpc away: $R_L = 1.5 \times 10^{20}$ m. The stars that astronomers are monitoring are either in the central bulge of our galaxy (10 kpc

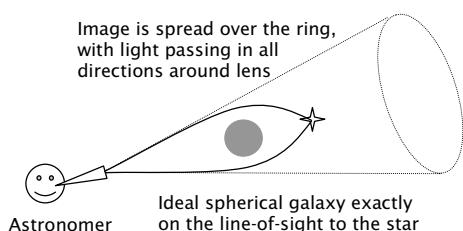


Figure 23.7. If the star, lens, and astronomer are lined up perfectly, and if the lens is a sphere, then the image will be a ring: light will reach the astronomer equally well traveling around the lens in any direction.

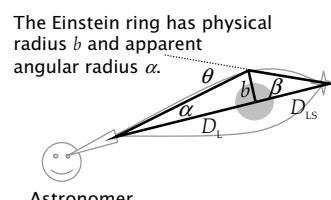


Figure 23.8. We can solve, to good accuracy, for the Einstein radius b (the impact parameter of the light) by working with the right triangles shown. See Equation 23.1 and Investigation 23.1 on the preceding page.

away) or in the Large Magellanic Cloud (50 kpc away), both of which offer large numbers of background stars in a small angular region. If we consider a hypothetic MACHO with a mass of $0.1M_{\odot}$, then the Einstein ring has an angular radius of about 1.4×10^{-9} radians, which translates to 0.3 milliarcseconds. (In linear size it is about 80 AU, about the size of the Solar System. This is much larger than we would expect the MACHO itself to be.) This is far too small an angle to resolve with telescopes, so one would not look for the deflection of the image due to the lens. Instead, one would look for the brightening caused by magnification if the Einstein ring of the lens happens to pass over the star. This phenomenon is called **microlensing**.

The mass of the lensing object can be estimated from the duration of the brightening event. The event lasts as long as it takes the Einstein ring to pass over the star, so if we can estimate the speed of the lens then we can calculate the size of the ring and hence the mass of the object. The velocity of the lens can't be inferred directly, but it can be estimated statistically by assuming that the population of MACHOS has a random velocity sufficient to keep them in a halo around the Galaxy, i.e. that they have roughly the circular orbital speed of an orbit at 5 kpc around an object with the mass of the Galaxy.

Recent results from two groups of astronomers who have constructed automatic, computer-controlled telescopes to do these repetitive observations are puzzling. They find MACHOS, but with masses that seem more consistent with $0.5M_{\odot}$ than with something below $0.1M_{\odot}$. On theoretical grounds, it is hard to understand how an object with as much as half a solar mass could not be a normal star, radiating with enough light to detect. But the lenses are definitely dark. Intensive follow-up observations with large telescopes have not shown any conventional stars where the lenses should be.

The number of detected MACHOS is not large enough yet for the statistics to be good enough to establish the case for the large-mass MACHOS beyond a doubt. But if an explanation based on lenses in the Large Magellanic Cloud does not stand the test of time, then these measurements have revolutionary potential. Physicists are already talking about possible new stable states of matter, boson stars, ways of transforming small white dwarfs into neutron stars or black holes, or something not yet thought of that will produce dark objects with half the mass of the Sun.

The third image: the ghost in a mirror

We are now ready to look at the third image that must be present if there is a second image of the type illustrated in Figure 23.6 on page 336. The existence of this image becomes evident if we ask what happens to rays of light that go *through* the lensing galaxy rather than around it. We shall assume for the moment that the light does go through without being absorbed or scattered. We are interested here in the effect of the geometry of spacetime on the propagation of light.

Because gravity is attractive, light rays will still be deflected in the same direction if they pass inside as if they passed outside. But the acceleration due to gravity begins to get smaller as one approaches the center of the galaxy: at the center it is exactly zero. So the amount of deflection is *smaller* for a ray passing near the center than it is for one further away. The effect of this is that the interior of the galaxy acts like a converging lens rather than an diverging lens.

This is illustrated in Figure 23.9 on the following page, which shows a situation in which the initial direct image is formed by rays that pass through the galaxy. The rays enter the galaxy diverging in the normal way. Their passage through the galaxy reduces their divergence. The image they form at the telescope therefore seems further away than the true distance of the source object.

The more interesting effect is what happens when there are two direct images,

►The calculation of the MACHO mass depends on assuming that the lensing MACHO is in the halo of our galaxy. But if in fact the lens is nearer to the star, say both in the Large Magellanic Cloud, then the mass required to produce the magnification goes down, because relative to the lens-star separation, the observer is much further away and will therefore require a much smaller deflection angle. So one possible explanation of this observation is that there are plenty of $0.1M_{\odot}$ objects in the Large Magellanic Cloud, but not many in our own Galaxy's halo.

In this section: we explain why smooth gravitational lenses always produce an odd number of images, why nearly half of them are mirror images, and why these are usually not seen in astronomical observations.

as in Figure 23.10. Here we see the same situation as earlier, with two direct images formed by a pair of galaxies. Let us first ask a simple question. If there are other stars in the field of view, besides the one we calculated the lensing for, where are they in the image? It is clear from the way the rays pass through the diverging (exterior) part of the lens that the ray coming from the star to the left of the main star arrives at the telescope also displaced to the left of the ray from the main star, and similarly for the star on the right. In other words, the left-right order of the sources is preserved in the image. This is what we mean when we say that an image is a **direct image**.

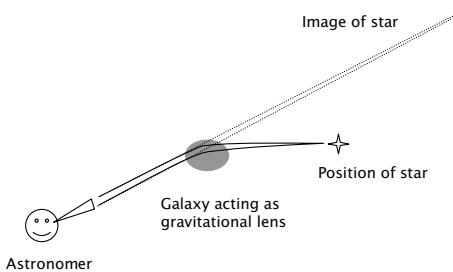
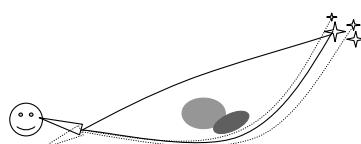


Figure 23.9. This is the same situation as in Figure 23.3 on page 333, but now the position of the galaxy is slightly different, so the lensed rays have to go through the galaxy to be seen by the astronomer. The ray closer to the center is deflected less than the ray further from the center, so the initial divergence of the rays is reduced. This puts the image location further away.

Now, it may seem that this is too obvious to spend time on. How could it be otherwise? The original image is also direct: a little experimenting with rays from the other stars will show that it also preserves the left-right ordering. And that is exactly the problem. To see why having two direct images creates a problem, consider sweeping the telescope slowly from the first image position to the second. This is a sweep to the right. For each position of the telescope, trace back the ray that enters it straight in. When the telescope points toward the first image, the ray that goes back to the star follows the upper curve. When the telescope points toward the second image, the ray goes back to the star again, this time taking the bottom path. At other positions, the ray will generally miss the star and go somewhere else. But what we will now show is that it has to hit the star again *somewhere in between the two direct image positions*.

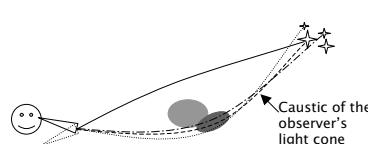
The matter distribution in the galaxies is continuous and smooth, so we should not expect any big jumps in the direction this ray takes far away. It should either pass to the left of the star or to the right of the star, but it won't jump from one side to the other without passing through the star.

So let us begin the sweep. The telescope begins by pointing at the first (main) image. As it sweeps to the right, the ray from the telescope encounters the star just to the right of the main star, because the first image is a direct image. This star appears in the telescope field of view. The sweep continues until the direction of the telescope is near the second image. Just before it points to the second image, the telescope is to the left of that position. As we can see from Figure 23.10, it will be



The astronomer scans the sky near one image.

Figure 23.10. When the astronomer looks at one of the images, she sees not just the central star but other stars as well if they are near enough. This figure illustrates the fact that a star to the left of the central star has an image that is also to the left, and similarly for a star on the right. The image is to the left because the diverging nature of the lens guarantees that the light from the star on the left will travel closer to the lens, and so it will appear in the image to have come from slightly to the left.



The astronomer looks at the third image, which goes through the converging part of the lens.

Figure 23.11. The third image must be formed by light that goes through the converging part of the lens, the mass of the galaxy. This figure shows how a ray from each of the stars reaches the telescope, pointing now at the third image. The rays cross, and the star on the left seems in the image to lie on the right. Because of the convergence, the third image also appears further away and dimmer (not illustrated in this diagram).

pointing at the star just to the left of the main star, again because the image is a direct one.

But how did this ray get to the left of the main star? After it left the neighborhood of the first image, it was pointing to the right of the main star. To get to the left of the main star again, *it had to pass through the main star somewhere in between the two images*. But if this ray passes through the main star, then the telescope sees the star in that position: there is a third image between the first two.

What is even more remarkable is that, as the telescope scans across the main star in this third image, *the image position is moving to the left as the telescope moves to the right!* This is, after all, how we deduced that there had to be a third image: we had to get the image position over to the left again. This happens as we sweep across the third image.

So the third image is inverted, with left turned to right. If the image were really of your grandmother cuddling her dog, and if she is right-handed, then in this image she is left-handed.

Where does this image form? The only possibility is that it forms in rays that pass through the galaxies. As long as the rays are on one side or the other of the galaxies, they will behave in the direct-image way as the telescope turns. What happens is that, in the galaxies, the convergence of the lens actually makes the rays cross, so that the ray from the left winds up on the right. The resulting caustic allows the inversion to happen. This is illustrated in Figure 23.11.

Because it is an image through a converging part of the lens, the third image seems further away, and hence dimmer. Moreover, in real galaxies there is a good chance that the light will be scattered or absorbed as it passes through dust clouds. So the third image will generally be hard to see. We should not be surprised if double images are more common than triple images in observations.

There is no reason to stop at three images. A really complicated lens could produce three or four direct images, maybe more. But an extension of the reasoning we have used here shows that, for every extra direct image, there must be another inverted image. It won't always be easy to detect, but it must in principle be there. As long as the gravitational field of the lensing galaxy is smooth, there will always be an odd number of images in principle.

You may have worried by now that we have only worked in a single plane: our diagrams and the reasoning from them have stayed in the plane defined by the telescope, galaxy, and star. But in reality, the galaxy will be an extended object with complicated structure out of the plane, so might that not have an effect? The answer is, of course, yes. Light rays can be deflected out of the plane, and image forming becomes more complicated.

For this reason, rather sophisticated computer programs are used to study realistic lenses and make maps of the image and lens structures. But the principles are the same as we have discussed here, and no fundamentally new kinds of images arise. In particular, there is a general mathematical theorem that the number of images must be odd, and that, after the first, they come in pairs: one direct, the other inverted.

The only exception to this theorem is if there is a gravitational field without a smooth galaxy or other smooth mass distribution. In Newtonian physics, this would be the field of a "point mass", a particle with zero size but finite mass. This is a mathematical device, but a fiction, so in all realistic Newtonian gravitational fields the theorem about an odd number of images holds. However, in relativity, black holes create a field with no smooth center. Lensing by a black hole does not need to create an odd number of images; the light that would form the odd images gets trapped by the hole instead of passing through to make the image. In fact, lensing

by black holes is a fascinating subject, since the light that passes near the horizon can circle the hole more than once before emerging. Images in such circumstances become very distorted!

Lensing shows us the true size of quasars

In this section: by using lensing one can measure the sizes of quasars, and they are astoundingly small.

►The astronomer Zwicky, whom we first met in Chapter 14, was the first to suggest, in 1937, that lensing by galaxies could be observable. As with his other work on neutron stars and dark matter, he was well ahead of his time.

Gravitational lensing has its main application in astronomy today in the field of cosmology. This is partly because of the numbers: if we take the equation for the Einstein radius, Equation 23.1 on page 337, and put in numbers that are appropriate for cosmological distances, then the angles look rather different. For example, suppose the lensing mass is a dense cluster of galaxies with $M = 10^{14} M_{\odot}$, at a distance of 100 Mpc, and suppose the lensed object is a quasar twice as far away. (This is nearer than typical lensing systems, but it will serve to illustrate the point.) Then the Einstein radius works out to be about 30 kpc, about three times the radius of our Galaxy. The angular size of this ring in an astronomer's telescope is about 1 minute of arc. This is much bigger than the microlensing rings, and it is an angle that modern telescopes can easily resolve, even from the ground. For this kind of lensing, finding images will be as important as finding brightening.

A picture of lensing, such as the image on the right-hand side in Figure 23.1 on page 332, can therefore yield quantitative information about the system. From the size of the Einstein radius, it is possible to estimate the mass of the lens. By comparing this mass to the mass that one would deduce from the brightness of the galaxies, one can estimate the amount of dark matter in galaxies and clusters. The amount found in this way is consistent with our other estimates, as described in Chapter 14.

The positions of the images are not the only information that astronomers can gather about the lens. If the lensed object is a quasar, it can be expected to vary in brightness by significant amounts on time-scales of days or weeks. These changes will not be seen at the same time in the different images, because the light rays of these images follow different paths. The paths have different lengths and, importantly, they experience different propagation time-delays due to the different gravitational redshifts they experience in the gravitational field of the lens. If one can guess the mass of the lens, say from photographs of the galaxies, then one can calculate the time-delay if one knows how far away the galaxies are. Since the time-delay is measured, one can use it to infer the distance to the lens. By measuring the redshift of the lensing galaxies, one can finally infer the value of the Hubble constant (Chapter 14). This is one of the key methods astronomers use today to measure the expansion rate of the Universe.

Alternatively, if one assumes a value for the Hubble constant, then one can use the measured time-delays to constrain models of the lens and pin down the overall mass and size of the lensing cluster. This is being done to determine the amount of dark matter in lenses, or indeed to discover regions in which there are condensations of dark matter with no visible galaxies at all.

One of the important side-effects of looking for correlated changes in brightness in the Einstein Cross lens system was the discovery that some brightness changes occur only in one image and not in the others. This does not mean that the images come from different sources, i.e. that the lensing model is wrong; the spectra of the four images are too similar for them to come from different objects. Rather, it indicates that another, short-lived lensing phenomenon may be acting. In particular, individual stars in the lensing galaxy will occasionally pass across the image and produce microlensing, just as we have described above. Scientists have shown that the number of observed events is consistent with what one would expect if the lensing galaxy has a population of stars similar to that of our Galaxy. More importantly,

the duration of a microlensing event is determined by the time it takes the lensing star to pass across the image, and this in turn depends on the size of the bundle of rays that form the image and the speed of the lensing star. Taking a speed consistent with a star orbiting within the lensing galaxy, scientists have calculated the size of the bundle of rays in the image, and they have then worked backward to calculate the size of the quasar itself.

Microlensing shows that the quasar in the Einstein Cross has a size no larger than 10^{13} m, or about 70 AU. This is an independent confirmation of the black hole model for quasars, since nothing else can be that small and yet have the required mass.

Weak lensing reveals strong gravity

As telescopes get larger and more sensitive, they can see and resolve the shapes of galaxies that are very far away. They can therefore also see the distortions in shape that are produced by intervening masses. While it is not possible to tell whether any single galaxy image has been distorted or just shows the true and irregular shape of the galaxy, it is possible to do this statistically if one can observe a large number of galaxy images in a single area of the sky.

The reason is that the distortions produced by a gravitational lens have a systematic orientation on the sky. Look again at the right-hand panel in Figure 23.1 on page 332. The distant lensed galaxy is always stretched in the direction along a circle surrounding the lens, and compressed in the radial direction. By contrast, any irregularities in the real shapes of galaxies should be randomly oriented on the sky. So if an image reveals slight distortions of the type seen in Figure 23.1, systematically arranged along circles surrounding a given center, then this is evidence of an intervening lens.

It might seem odd that images get compressed radially and stretched along the circle, since a gravitational lens is a diverging lens, and especially if we recall Figure 5.1 on page 40, which shows that the trajectories of falling objects are stretched apart in the radial direction and squeezed together in the horizontal direction. The situation is no different for light, but the effect of this is to compress the *image* radially, not stretch it. This is clear from Figure 23.4 on page 334, which displays the effect only of the radial compression of the geodesics. The figure shows that the light arriving into a given angle at the telescope covers more of the star than if the lens had not been there. This means that without the lens the image of the star would be larger in this direction. In the horizontal direction the effect is just the opposite.

As this chapter is being written (2002), astronomers are just beginning to employ the power of this statistical method to make estimates of the masses of the dark matter halos of clusters of galaxies. These estimates appear to be consistent with other measures of the dark matter, which is reassuring. Over the next few years, astronomers may be able to produce a complete map of the dark matter distribution in the nearby Universe. This is one more tool for discovering the way galaxies formed as the Universe expanded from its initial singularity. It is time, therefore, for us to turn our attention to the Universe as a whole, and begin to study cosmology.

In this section: by studying the very small distortions of hundreds of images, astronomers are beginning to detect the clumpiness of the dark matter distribution in the Universe.

Cosmology: the study of *everything*

Cosmology is the study of the Universe as a whole. A century ago, scientists had only a vague idea about what even the Milky Way galaxy was like, and they were only able to make guesses about the Universe beyond. Most educated people believed that the nature and history of the Universe were simply matters for religious belief. The word “cosmology” referred to the set of beliefs one had about the whole world: Earth, God, Universe, Creation.

It is one of the most remarkable achievements of modern astronomy that it has turned cosmology into a scientific discipline. In fact, cosmology is one of the most active and productive areas of scientific research today. Stunning pictures taken by the world’s most powerful telescopes have shown us what the Universe looks like at very great distances and very early times (Figure 24.5 on page 353), and they have allowed us glimpses of its very early history (Figure 24.6 on page 354). Indeed, as we shall see below, they have brought us the most startling revelation of all: the Universe had a beginning.

Every year new observations bring a greater understanding of how the Universe began and how it evolved, of where galaxies and stars came from and how they led to the evolution of life. Scientists are even beginning to trace changes in the very laws that govern the behavior of the elementary particles, from the beginning of time to the present day. And woven through it all, regulating every step, is gravity.

One of the things that makes the science of cosmology so very interesting is its tendency to stir up controversy. Because the “study of everything” brings astronomy into areas that have been the preserve of philosophy and religion, many questions in cosmology arouse emotional debates that are far more intense than in any other branch of physics.

The discovery that there appears to have been a beginning to time itself – the Big Bang – has made many people question what their religions teach them about Eternity and Creation. Some scientists try to build bridges between religion and modern cosmology, but this turns out to be a personal exercise, and one scientist’s explanation may offer little comfort to another. Some non-scientists seem to have felt so threatened by the notion of the Big Bang that they have rejected the validity of the scientific approach itself.

So cosmology is important to everyday life in a way that other interesting subjects in astronomy, such as the study of the giant black holes in the centers of galaxies, are not. Although its practical applications and commercial “spinoffs” are negligible (see Figure 24.1 on the next page), cosmology has nevertheless become one of the most important branches of modern physics.

Cosmologists believe that they can make progress toward understanding deep cosmological questions – such as the Big Bang and the nature of time – in a scientific

In this chapter: we introduce our study of cosmology. We focus on the measurements that astronomers can make about the Universe as a whole: the Hubble expansion and the acceleration of the Universe. We learn about homogeneity and the Copernican principle, about what the expansion does to space and what is in it, and how to compute the evolution of the Universe.

►The image under the text on this page reminds us that creation myths and cosmologies were central parts of the belief systems of ancient peoples. It is remarkable how many cultures believed in a beginning of time, a moment of creation. The ancient Egyptians had several creation myths. The Hebrew creation story even orders the events in much the same way that modern science would, although on a vastly different time-scale. More than any other branch of physics, the scientific study of cosmology raises religious sensitivities and addresses questions that have long been regarded the domain of philosophy and belief. Image from a photograph copyright Jon Peck, www.ancient-mysteries.com, used with permission.

way, by the same methods that scientists use when trying to discover what is going on in the centers of galaxies. Like any scientific study, cosmology cannot answer every question, and indeed the understanding of many issues is very incomplete. This leaves plenty of room for religious belief and lively philosophical debate. Also like any other science, cosmology almost inevitably suggests new and exciting ideas, such as the possibility that by studying the Universe as a whole physicists can learn about the laws that govern the innermost workings of protons and electrons. And sometimes the new ideas raise troubling questions, such as what came “before” the Big Bang, and even whether the notion of “before” can make any sense.



Figure 24.1. “My big mistake was going into cosmology just for the money.”

This is the wrong way to think about cosmology! Copyright by S Harris, used with permission.

In this section: cosmology studies the observable part of the Universe, inside the *particle horizon*.

▷ The particle horizon is very different from the *event horizon* of a black hole. The event horizon prevents us from learning about what is inside, even in the future. The particle horizon simply tells us how far we can see at the present moment.

I will return to these issues in the last of our four cosmology chapters, Chapter 27, after we have studied the scientific foundations of cosmology in astronomical observations and in the theory of general relativity. These four chapters on cosmology will have a different character from earlier chapters. Here we confront the limits of our knowledge, the limits of what our theoretical understanding of physics tells us. Studies of cosmology are even today calling fundamental assumptions into question and revealing places where new physical theories are needed. These chapters are a snapshot, at a time of rapid change, of a field in

which some of the deepest questions in physics are being asked and at least partially answered. Cosmology is at the sharp end of physics, and inevitably some of what we discuss here is uncertain, tentative, speculative, paradoxical ... and wonderfully exciting!

What is “everything”?

Some cosmologists like to think that they are studying the whole Universe, possibly infinite in extent, from the beginning of time to the present. They will tell you that the entire Universe is like this or like that. Don’t believe them! There are parts of the Universe that are too far away to see, and all we can do is speculate about them. Let us start our study by being careful about what it is we can and cannot investigate.

As in any other scientific study, we are only able to describe what we can observe, and in cosmology we can only study the part of the Universe that is *near enough* for us to observe. If we accept for the moment that the Universe began a finite amount of time ago, then there is a limit on how far away we can see anything. The most distant parts of the Universe which we can have any hope at all of observing are those regions that, at the time of the Big Bang, emitted photons or gravitational waves which are just reaching us today, having traveled at the speed of light ever since the Big Bang. The past light-cone that stretches backwards in time from our present moment is the boundary of this region, and it is called our **particle horizon**, or simply our horizon. This is illustrated in Figure 24.2. Any region further away (outside the particle horizon) has not yet had time to send us any signals, so we cannot even in principle say anything scientific about it at all.

This is then the province of *scientific* cosmology: the observable Uni-

verse. Its size increases daily: every new observation could in principle bring information from a region of the Universe that until that moment was too far away to see. In practice, astronomical instruments are not yet good enough to see to these extreme theoretical limits, but in the next couple of decades gravitational wave detectors might achieve this. (See Chapter 22.)

Importantly, it is always possible that new information from a previously unobserved part of the Universe could upset all of our old theories. Tomorrow's observation might reveal a more distant region of the Universe which is, say, infinitely old, which did not experience a Big Bang and a beginning of time. Cosmology is just like all other sciences: new data can destroy old theories.

But cosmologists have a reason for not expecting anything as radical as this to happen to their picture of the Big Bang. They call this reason the **Copernican principle**, or more plainly the **principle of mediocrity**, and it can be tested by observations.

Copernican principle: "everything" is the same "everywhere"

Copernicus simplified our picture of the planets by dropping the assumption that the Earth was specially located at the center of the Solar System. He realized that the Earth was just one of the planets moving around the Sun. Later scientists came to understand that the Sun is just an ordinary star, moving about the center of our Galaxy, and that the Milky Way is a fairly ordinary galaxy, a member of an unremarkable small cluster of galaxies called the Local Group.

We are, in effect, *mediocre*: we are not at the center of anything, and anyone looking at the Universe from a different location would not immediately be drawn to look at us by any special features of our neighborhood. What we see around us is typical of the Universe as a whole, and it seems reasonable to assume that this is true even of its unseen parts. This assumption is called the *Copernican principle*.

This does not mean that every part of the Universe is as boring as our immediate neighborhood; far from it! The Universe contains quasars, active galaxies, gamma-ray bursters, black holes, and a host of other exotic objects. But there are no more of them anywhere else in the Universe at the present time than there are around us.

The Copernican principle also reassures us that we shouldn't live in fear (or excited anticipation!) of seeing a completely different kind of Universe crossing over our particle horizon the next time we look in our telescopes. If this were to happen, then our moment in *time* would also be very special. This is because we can already see enough of our Universe to know that hypothetical cosmologists at some earlier time (say, a billion years ago on another planet elsewhere in the Milky Way) would have had no chance of seeing big changes to their picture of the Universe in short times.

These expectations are *not* scientific deductions. The Copernican principle is basically a philosophical assumption, but this does not make it unscientific. Scientists often use such assumptions as important guides to the framing of scientific theories, helping them to choose among a multitude of possible avenues to explore. The Copernican principle has the great advantage of simplicity: it tells us that we don't need to waste endless time speculating about what distant regions of the Universe look like. Unless we are forced by new evidence to think otherwise – and no philosophical principle should blind us to hard evidence – we shall assume they look the

In this section: the observable Universe is homogeneous on large scales. Our location in space is not particularly special.

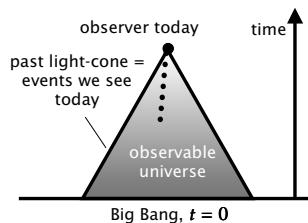


Figure 24.2. The part of the Universe that we can observe, at least in principle, is limited to the regions that have had time to send us light since the Big Bang. The past light-cone of the observer is the boundary of the observable region, called the particle horizon. Tomorrow, when the observer makes another observation, this boundary will have moved upwards and outwards. So the region outside the boundary contains events that could someday become observable. Using light, the observer sees only the events on the light-cone, but the events inside could have sent information at slower speeds. The observer's own history (that of the Earth and the elements from which it was made) is shown as the dotted line. This diagram over-simplifies the structure of the Big Bang itself, but it conveys the correct idea about the boundary of the observable region.

Investigation 24.1. The rubber-band universe

Here we look in some detail at how the rubber-band universe behaves, how it produces a Hubble law for people in it. The main difference between this “toy” universe and the real one is dimension: the rubber band is one-dimensional, whereas the real world is three-dimensional. But if we look in a fixed direction in our three-dimensional world, we will see a one-dimensional strip that behaves just like our rubber band.

The first point, and the most crucial, is that the rubber band is the whole universe to its inhabitants. They cannot leave it, nor can light signals or any other kind of physical means of transporting information. When one dot on the band looks at another dot, it looks along the band, not in a short-cut across it. That would take it into another dimension, outside of its one-dimensional universe. While this is not a hard rule to visualize in the case of a rubber band, it applies equally strongly to us in our real Universe. When we come to discuss the geometry of possible universe models in Chapter 26, it is important to bear in mind that we can in principle only measure within the three dimensions of the space, not into any extra dimensions that we may have to use for visualizing the model.

Now suppose that the circumference C of the band behaves in some arbitrary way as a function of time, denoted by $C(t)$. Then all separations between dots will change in exact proportion. So if dots 1 and 2 are separated by a distance $d_{1,2}(0)$ at time $t = 0$ (when the rubber band is relaxed), then at any other time t their separation will be

$$d_{1,2}(t) = \left(\frac{C(t)}{C(0)} \right) d_{1,2}(0).$$

Let us call the ratio of circumferences the *scale-factor* $R(t)$ of this universe:

$$R(t) = \left(\frac{C(t)}{C(0)} \right).$$

Then the equation governing the separation of dots along the band is

$$d_{1,2}(t) = R(t) d_{1,2}(0). \quad (24.1)$$

Now, the average speed of separation of the dots, $v_{1,2}$, as measured along the band by either of them, will be the change of distance, $d_{1,2}(t) - d_{1,2}(0)$, divided by the time t . This gives

$$v_{1,2} = \frac{d_{1,2}(t) - d_{1,2}(0)}{t} = \frac{R(t) - 1}{t} d_{1,2}(0). \quad (24.2)$$

We have derived Hubble’s law here: the speed of separation is proportional to the original separation $d_{1,2}(0)$. The rubber band is a good model of an expanding or contracting Universe.

Exercise 24.1.1: Age of the Universe

Assume that the Hubble constant has a value $70 \text{ km s}^{-1} \text{ Mpc}^{-1}$. First convert this value to more standard units (s^{-1}) by converting megaparsecs to kilometers. Then take its reciprocal to find that the approximate age of the Universe is 14 billion years. If the Hubble constant is larger, what does this do to the approximate age?

It is important that we have not imposed any particular stretching law: $R(t)$ is an arbitrary function of time. But we have insisted that the change of length takes place uniformly along the band, so that the relative distances remain the same. This is the equivalent of assuming the Universe is homogeneous.

Now suppose that the change is actually a contraction at constant speed v , so that

$$C(t) = C(0) - vt. \quad (24.3)$$

Then the rubber band will contract to a point (reach zero circumference) at a time

$$T = C(0)/v.$$

For this contraction law, the scale-factor is $R(t) = 1 - [vt/C(0)]$, and the Hubble law is (from Equation 24.2)

$$v_{1,2} = -v \frac{d_{1,2}(0)}{C(0)}.$$

This is reasonable: the speed of approach of any two dots is the same fraction of the overall contraction rate v as their separation is of the whole circumference. In this case, the Hubble constant we infer is

$$H = -v/C(0),$$

which, not surprisingly, is constant for all time. The key result we have been looking for is that this is just the reciprocal of the time to contract to zero (apart from a sign):

$$T = |1/H|. \quad (24.4)$$

Now, the same formula gives the age of a uniformly expanding rubber band as well. If the rubber band started at zero circumference and expanded, then a movie of it run backwards would show a rubber band contracting to zero circumference with constant speed. The age of the original expanding band is the same as the time it takes the contracting band in the movie to reach zero size. We have therefore learned that if a rubber band (or a universe) expands at a constant rate, then the reciprocal of the Hubble constant gives its age. We show in Exercise 24.1.1 that the approximate age of the Universe is 14 billion years.

Of course, because of gravity, the expansion of the Universe is not constant: it may slow down or even speed up, as we will see later. That means that the reciprocal of the Hubble constant is only an approximation to the age. The Hubble constant is a constant in space if the Universe is homogeneous, but it is not constant in time.

same as the region we can see. This idea is related to Occam’s razor, which we met in Chapter 19. The importance of simplicity is deeply ingrained in physicists’ thinking.

The Copernican principle must be tested, of course: we cannot accept any guiding principle if it predicts things that conflict with observations. It is supported by a number of remarkable observations. Everywhere we look in the Universe, we observe that its appearance in one region is very similar to that in another with the same cosmological age. Astronomers have measured properties like the number of galaxies in a given volume; the shapes and colors of galaxies in different places; the speed of the Hubble expansion (see Chapter 14); and even some of the most fundamental numbers in physics, such as the ratio of the mass of the proton to the mass of the electron. In each case, the properties are the same everywhere they can be measured. This fact has a name: the Universe is *homogeneous*.

On the other hand, modern observations show that the Copernican principle does not hold in the time direction: the Universe was decidedly different long ago, filled with quasars expelling enormous jets of gas, with early generations of very blue and very hot stars, with gas poor in the heavier elements that are needed for life, with galaxies much nearer to one another than they are today, and hence with many colliding and merging galaxies. Today, the expansion of the Universe has turned it into a quieter place. We shall now see how we can understand this expansion.

The Hubble expansion and the Big Bang

We saw in Chapter 14 that Hubble measured the recession speeds of galaxies, and found that the speed of a galaxy was proportional to its distance from us:

$$v = Hd. \quad (24.5)$$

What does this mean about the Universe? Can we understand the Hubble law in a simple way? The answer is yes.

There is a simple way to demonstrate how Hubble's law arises in an expanding Universe. Let a rubber band represent a one-dimensional "universe": only points actually on the band represent points in this one-dimensional universe. Draw dots on it to represent galaxies, as in Figure 24.3. Arrange it as a circle, and then stretch it to two, three, and four times its original radius, retaining its circular shape as far as possible. Then the length of the arc of the circle between points separated by, say, a quarter of a circle increases at a uniform rate, which is half the rate that the length of the arc joining points separated by a half-circle does. The more distant are any two points, the more rapidly they move apart. This is just Hubble's law.

The rubber-band universe is explored more quantitatively in Investigation 24.1.

Notice that the rubber-band universe really is homogeneous: no dot occupies a special position, there is no natural "central" dot on the band. (The circle does have a center, but that is not part of the band. In our one-dimensional universe, only points along the band are part of the universe.) Because every dot is like every other one, all dots see the same Hubble law. Every observer attached to a dot sees the universe expanding away from the "home" dot. This is a good model for what our Universe looks like, except we have to extend the model to three dimensions.

Did our rubber-band universe have to be a closed loop? No: if it were a long straight piece of rubber and we drew dots on it and then stretched it, we would have seen the same thing. The Hubble law would still apply, as measured by an observer on any dot. And still no dot would be in a special position.

Cosmologists make a distinction between *local* and *global* properties of the universe. Local properties are measurable directly. Global ones are properties of the Universe as a whole; as we emphasized earlier, these are usually more hypothetical. The Hubble expansion, for example, does not define the *global* structure of our rubber-band universe: it cannot tell us whether the Universe looks like a straight piece of rubber or a circular rubber band. It only tells us how it stretches, locally. In Chapter 25 and Chapter 26 we will take up the subject of what the Universe looks like in the large.

In this section: we see how a homogeneous expansion of a homogeneous space leads automatically to the Hubble law.

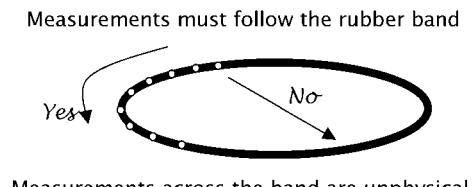


Figure 24.3. The rubber-band model of the Universe.

Our picture of the Universe as a homogeneous, expanding “gas” of galaxies is really very simple. What does this tell us about the Universe at earlier times?

If we add to the picture our expectation that gravity is attractive, so that the different parts of the Universe will be pulling back on each other, then we expect that the expansion should be slowing down. If so, then the expansion rate in the past would actually have been *faster* than it is today, and at some finite time in the past all the gas and galaxies in the Universe would have been squeezed together to an infinitely high density. This moment of infinite density is another example of a *singularity*, just as we found inside a black hole. Where the singularity in the black hole is the end of time for any particle that encounters it, the cosmological singularity is the *beginning* of time for all particles in the Universe. The expansion of the Universe away from this singularity is what we call the Big Bang.

- ▷ The homogeneity of the Universe makes this infinite density inevitable: everything in the Universe came together at exactly the same moment.

In this section: astronomers have discovered that the expansion of the Universe is getting faster with time.

We look at their methods, and especially the evidence from the supernovae that they use as standard candles to measure distances.

The accelerating Universe

However, our simple and apparently obvious assumption that gravity is attractive has recently been called into question by astronomical observations. Astronomers have been able to use automated techniques to find large numbers of very distant Type Ia supernovae, and their measurements are providing a strong indication that the expansion of the Universe is actually accelerating today, that the Hubble constant was smaller in the recent past than it is today. If this is the case, then a singularity at the beginning of time is not an inevitable consequence of physical laws. It is a matter for observation, measurement, and physical theory to decide if it really took place.

We have often discussed supernovae of Type II in this book: they are triggered by the collapse of the core of a giant star, and they result in the formation of a neutron star or a black hole. But not all stellar explosions are triggered in this way. Another way to make an explosion is by accreting sufficient matter onto a white dwarf.

If a white dwarf star finds itself in a binary system with a star that is shedding matter, then some of that matter will accrete onto the white dwarf, just as in binary systems containing neutron stars or black holes. If accretion goes on long enough at a high enough rate, then the mass of the dwarf will increase to the Chandrasekhar mass (Chapter 12), the maximum mass allowed for the dwarf. When this happens, the star must collapse.

But the star does not usually collapse to a neutron star. Instead, if the nuclei in the white dwarf are relatively light, not having been processed as far as iron, then nuclear reactions begin to take place during the collapse. The star becomes an enormous nuclear bomb. The energy released in the runaway reactions disintegrates the star, and the result is a huge stellar explosion, a different kind of supernova.

The interesting thing about such explosions, from the point of view of measuring the cosmological constant, is that they are not very dependent on how the white dwarf was formed or on what kind of companion star is shedding material onto the dwarf. The Chandrasekhar mass depends only on fundamental constants of physics, so it should have been the same for the first stars as it is today. This mass largely determines the properties of the explosion. Within tight bounds, there is therefore not much variation in the explosion from one event to the next: the peak luminosity, for example, is fairly constant. Astronomers call such a system a “standard candle”, a term we introduced in Chapter 9.

Here is how astronomers use Type Ia supernovae to determine the value of the cosmological constant. When a supernova in a very distant galaxy is detected, it is observed carefully, night after night, in order to find the maximum brightness. Since astronomers already know the intrinsic luminosity of the supernova at maximum

brightness, they can then compute how far away the source is, and in particular what its redshift should be if we assume that the Universe has a particular Hubble constant and deceleration parameter.

The astronomers then directly measure the redshift of the galaxy containing the supernova and compare this with the predicted redshift. They adjust the Hubble constant and deceleration parameter until the predicted redshift matches the observed redshift, and in this way they determine these parameters.

This can only be done in a statistical way, of course. All measurements have some uncertainties, and so to get a good result astronomers use many tens of supernovae. These are searched for using automated telescopes that survey small regions of the sky to look for very dim and distant galaxies. When the brightness of a galaxy changes, the astronomers must investigate further to determine if it was a supernova, and then if it was of Type Ia.

At the time of writing (2002), two independent groups of astronomers had measured these parameters. The data of the two teams are illustrated in Figure 24.4 on the next page. The value of the Hubble constant indicated here is near to other estimates, around $70 \text{ km s}^{-1} \text{ Mpc}^{-1}$. But the deceleration parameter they measure is negative: the expansion of the Universe is accelerating!

These observations are very recent, and not all alternative explanations of the observations have been fully explored and ruled out. The scatter of the points in Figure 24.4 is large, and this raises the possibility that a small systematic effect, not so far taken into account, could change the conclusion, but intense work on this question has so far not found any reason to doubt the result. More data are needed, and astronomers are building new instruments and even satellites to gather it. But for now, the evidence for acceleration is strong, and it has led to a thorough re-examination of the physics and astrophysics of cosmology. Even if the acceleration proves in the end to be an illusion, the stimulus it has given to physicists and astronomers to come up with new ideas and justify old ones has been a positive result.

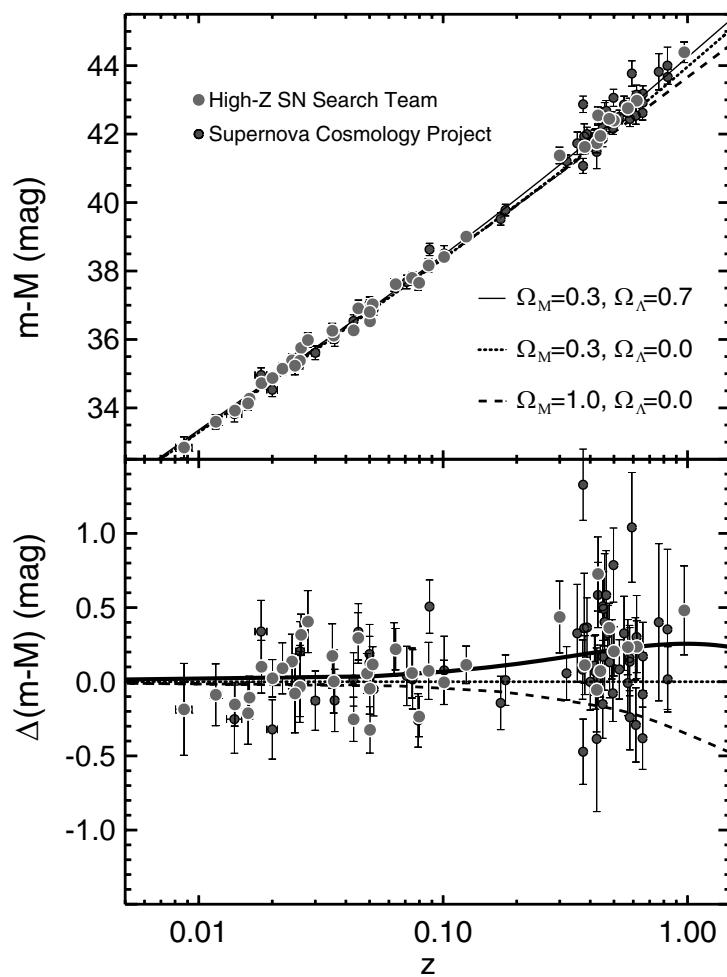
Was there a Big Bang?

The acceleration of the Universe was unexpected to most physicists, and there is as yet no accepted theoretical explanation of it. Evidently the Universe is filled with a physical field that exerts a repulsive gravitational effect that is larger than the attraction of all the normal matter in the Universe. We saw in Chapter 19 that it is possible to make the active gravitational mass negative, i.e. to achieve repulsive gravity, if the physical field has a large negative pressure. Most explanations of the accelerating Universe propose some such field. Einstein's cosmological constant, introduced in Chapter 19, is such a field. It may well be the explanation of the acceleration, but there are other alternatives. Astronomers are beginning to call this field **dark energy**, to draw a parallel with dark matter. Just like dark matter, dark energy is invisible except through its gravitational effects. But unlike dark matter, dark energy creates anti-gravity effects. We will see in Chapter 27 that it is likely that the eventual explanation of dark energy will have deep implications for theoretical physics.

Even without a theory we can ask what implications the observation of acceleration has for the Big Bang. Going backwards in time, the normal matter of the Universe gets denser, and so its contribution to the net force of gravity increases. If the dark energy behaves like a cosmological constant, then (as we saw in Chapter 19) its repulsive effect will remain roughly constant in strength. Given the fact that the observed acceleration is of roughly the same size as the expected *deceleration*, so that the gravitational effect of the dark energy is comparable to that of the matter

In this section: the Big Bang must have occurred if gravity was always attractive in the past; but dark energy might have prevented it.

Figure 24.4. Trend of distance against redshift for distant supernovae of Type Ia. This figure shows the evidence for the acceleration of the Universe at the time this book was completed (2002). In the upper plot, the horizontal axis gives the redshift of the supernova, as inferred from its spectrum and that of the galaxy containing it. This measures the speed of the expansion of the Universe (Chapter 14). The vertical axis gives the difference between the apparent magnitude m and the (standard) absolute magnitude M of the supernova, which (see Chapter 9) is a logarithmic measure of the distance to the object. In a simple cosmology with constant expansion speed (no acceleration or deceleration), the points would all fall on a straight line. Accordingly, the lower plot shows only the deviation of the points from that straight line. The expected relations for three model universes are drawn as lines. Don't worry at this point about the notations Ω_Λ and Ω_M , which we will define in later chapters. The line of long dashes is a universe that is decelerating and whose speed will approach zero as it gets larger and larger; the line of short dashes is a model that is decelerating less rapidly, so that it eventually will expand at a constant speed; and the solid line is a Universe that is accelerating. The points have a large scatter because of observational uncertainties, but the solid line gives the best fit: the deviation of the points from this line is significantly less than that from the other lines. The data come from two independent teams of astronomers, measuring different supernovae. Figure from High-Z Supernova Search, based on data from Riess, A. G., et al (1998) *Astronomical Journal* **116**, 1009, and Perlmutter, S., et al (1999) *Astrophysical Journal* **517**, 565.



today, it follows that when the Universe was even one-half of its present age, the attractive (matter-created) part of gravity dominated, and the Universe was slowing down. On this hypothesis, the Universe had attractive gravity as long as it was smaller than this.

The net attractiveness of gravity at early times would still have made the Big Bang unavoidable. If we take our rubber-band universe and follow it back in time, assuming that it was always expanding at a constant rate, then at a *finite* time in the past, it was a single point. We show in Investigation 24.1 on page 348 that this time is just the reciprocal of the Hubble constant: if the speed increases with distance as $v = Hd$, then a universe with a constant expansion rate would have begun as a point at a time $T = 1/H$ ago. This is called the **Hubble time**, and in Investigation 24.1 on page 348 we show that its value, as measured by astronomers today, is around 14 billion years. If the net effect of gravity was actually attractive at early times, then the universe would have been expanding faster in the past than now, so that it would have actually taken less time to arrive at its present size than if the expansion had been constant. Therefore, the time $1/H$ would in fact be an *overestimate* of the age of the universe.

However, if the dark energy could have increased in density for some reason in the past, so that it always maintains its advantage over the matter (including dark matter), then the Big Bang is not inevitable. The Universe might have started its expansion from rest at a small but finite size, propelled to expand by the repulsive field. Since we don't know the properties of this repulsive field, we can't use theory alone to decide which of these two alternatives really happened. Instead, we have to appeal to observations.

We will see in this chapter that the observations strongly favor the Big Bang, so they suggest that the dark energy was weak at early times. However, no matter how far back in time our observations allow us to go in testing the existence of the Big Bang, it would still be possible to postulate that the repulsion at times earlier than we can directly observe was much stronger than at later times.

Indeed, the concept of inflation (Chapter 27) relies on just such a repulsion at very early times, and it is possible that inflation itself actually reversed an earlier contraction phase, preventing a true singularity. In this case, the Universe would have been small enough and dense enough to have looked like the conventional Big Bang in all of its measurable consequences, but it would have avoided a singularity.

In such theories (which seem unlikely to at present) one needs to distinguish the Big Bang, which refers to the expansion from a hot, dense state, from the singularity, which is the very beginning of time. Even if the singularity did not occur, the hot dense Big Bang is necessary to explain the formation of elements, of the cosmic microwave background, and of galaxies, all of which we study in the next chapter.

Only further research will tell what the nature of the dark energy is. It is very possible that the answers will come from a future theory of quantum gravity, which we will consider in the final chapter. Meanwhile, we will focus on today's "standard" model of cosmology, in which the repulsive field we see today was not important at early times, and the Big Bang is the expansion of the Universe away from a genuine singularity.

Looking back nearly to the beginning

The most remarkable tests of the Copernican principle are a number of observations that look at things that are very far away indeed, and therefore very far in the past. Consider, for example, the galaxies in Figure 24.5. The light reaching us now from those objects has been traveling toward us for more than 80% of the time since the Big Bang. When astronomers look at the most distant galaxies in any particular direction, they are looking back in time most of the way to the Big Bang.

Now suppose we look at two groups of very distant galaxies that are in opposite directions from us. Light from a galaxy in one group has taken 80% of the age of the Universe to reach us; light has surely not had enough time to reach the galaxies in the other group that we can see in the opposite direction. That means that the galaxies in one group have not yet entered the observable universe of the other group.

>Interestingly, the observed near-balance between repulsion and attraction today means that we apparently live in a special time in the history of the Universe. At much earlier times, the repulsion of the dark energy was not noticeable. At much later times, the attractive pull of the matter will probably also be negligible. It seems that the Universe has two comparable effects nearly in balance only at roughly the stage in the history of the Universe where people are measuring things. This is very non-Copernican. It seems hard to apply the Copernican principle at all to our place in time, even if it applies to our place in space.

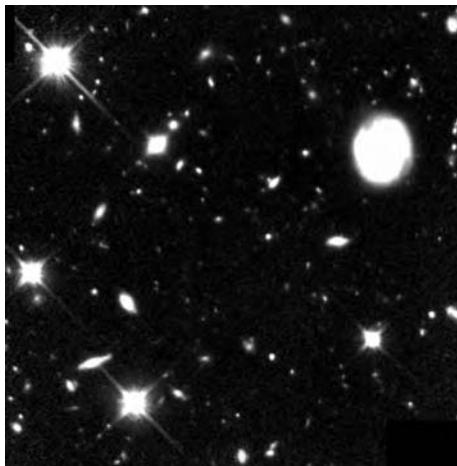


Figure 24.5. The Hubble Deep Field South is one of the longest-exposure pictures ever taken with the Hubble Space Telescope. It reveals galaxies as they looked when the Universe was only one-fifth of its present age. A photo of the North field, in the opposite direction, shows very similar galaxy forms and numbers, even though there has not been enough time since the Big Bang for the more distant regions to have affected each other in any way. Photo courtesy HDF-S Team and NASA/STSCI.

In this section: the Big Bang seems to have started in the same way everywhere we can see.

Intelligent beings living today near the site of one group of galaxies can see our locality (as it was when the Universe was only 20% of its present age), but they can not yet see the galaxies in the other group. Yet we can see both groups, and we can see that the Universe around one group looks very similar indeed to the Universe around the other. So if people today living near the site of one galaxy are assuming the Copernican principle about the unseen region of the Universe in the direction of the other set of galaxies, then we can say that they are right.

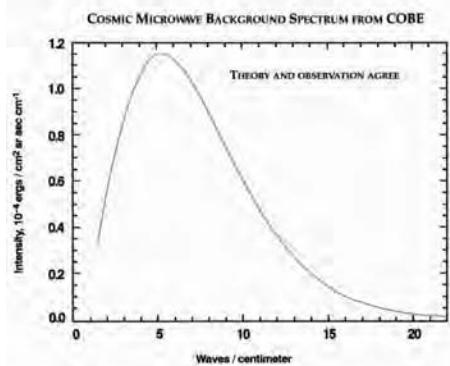
Even more striking support for the Copernican principle comes from observations of the cosmic microwave background radiation. We turn to a discussion of this, which is the most important piece of evidence for the Big Bang at this time.

Cosmic microwave background: echo of the Big Bang

The early Universe was unimaginably dense and hot. Ordinary matter as we know it did not exist. At early enough times the temperature was so high that matter formed a plasma: the atoms of the gas were moving so fast at this temperature that when they collided they ionized one another, stripping the electrons away from the nuclei.

In this section: our earliest direct signal from the Big Bang is the microwave background. We learn how it was accidentally discovered and why it has such fundamental importance.

Figure 24.6. The spectrum of the cosmic microwave background radiation, as measured by the COBE spacecraft. It follows the expected black-body curve perfectly. The uncertainties in the measurements are smaller than the width of the line. This radiation was emitted when the Universe was only 3×10^5 years old, a fraction 2×10^{-5} of its present age. Courtesy COBE team and NASA/GSFC.



scatter frequently from the charged particles and quickly come to have the same average energy. They form a photon gas.

Now, dense plasmas are good “black bodies”, as described in Chapter 10: any radiation that falls onto such a plasma will instantly be scattered by the charged particles and trapped within the plasma. It will not get through to the other side or be reflected. Because of Planck’s remarkable law of black bodies, the radiation in a dense plasma will have a black-body spectrum. This was the case when the Universe was young and hot.

As the Universe expanded, its plasma cooled, and eventually the matter became cool enough for the electrons to combine with the nuclei to form a neutral gas. At this point, the radiation in the plasma suddenly became free. Photons prefer to scatter from charged particles, not neutral atoms, so once the particles in the Universe had become neutral, the typical photon would never be scattered again. For this reason the moment when electrons and nuclei combine is called the epoch of **decoupling**, when photons effectively decoupled from the rest of the matter.

At that moment, the photons had the spectrum of a black body with a temperature hot enough to ionize hydrogen. As the Universe expanded further, this photon gas also expanded, and like any expanding gas it cooled off. Since these photons

►The epoch of decoupling is also frequently called “recombination”, to indicate what is happening to the electrons and protons. However, this is an odd name for this event, since at earlier times the Universe was so hot that electrons and nuclei had never previously been combined. A better name for this epoch would be “first combination”, but it seems too late to change this usage. Along with many other physicists, I prefer to use the term *decoupling*.

have been moving freely through the Universe ever since, not scattering from anything, they have experienced a cosmological redshift, just as the light emitted later by galaxies has. (See Chapter 14.) This redshift applies to every photon, so the net result is that the photons still have a black-body spectrum, but its temperature has been redshifted to a much lower value. The details of how this happened are studied in Investigation 24.2 on the next page.

These photons have been observed, and they are in the microwave part of the spectrum. Their temperature, about 2.7 K (Figure 24.6), is a factor of about 1000 lower than that of a plasma that can ionize hydrogen. Investigation 24.2 on the next page shows us that this implies that the Universe has expanded by the same factor of 1000 since decoupling. Allowing for the overall slowing of the Hubble expansion, the Universe was about 0.003% of its present age when this radiation was emitted, roughly 300 000 y old.

The cosmic microwave background radiation was discovered by accident by the American physicists Arno A Penzias (b. 1933) and Robert W Wilson (b. 1936) in 1965 at Bell Laboratories, because it was an unexpected and annoying source of noise in microwave transmissions of telephone conversations! For recognizing the importance of their discovery, Penzias and Wilson received the Nobel Prize for Physics in 1978.

Since the plasma of the Universe scattered photons easily before decoupling, the Universe was essentially opaque to photons then. Therefore, we have no hope of ever detecting any electromagnetic radiation that originated at a time earlier than the cosmic microwave background.

When we look at the microwave background, we are seeing the Universe at a very early age. It is very significant, therefore, that the microwave background has the same property as galaxies: when we look at it in one direction and in the opposite direction, we see exactly the same sort of radiation, with the same temperature and the same degree of uniformity.

In fact, we don't need to look in opposite directions: because the radiation originates so much closer to the time of the Big Bang than does the light given off by galaxies, the microwave radiation coming to us from directions separated by only a couple of degrees on the sky is coming from regions that would have had no knowledge of one another at the time the radiation was emitted.

The rest frame of the Universe

Because it brings us our earliest glimpse of the Universe, the cosmic microwave background has been studied intensively by astronomers. It has a number of lessons to teach us, and they all suggest that the Copernican principle is correct to a very high degree of accuracy. One of the most remarkable is that it determines the mean rest frame of the Universe.

The microwave background does this because it enables us to measure our own speed relative to this rest frame. In each direction that astronomers look, the spectrum of the cosmic microwave background is a black-body spectrum, characterized by a single temperature. If the Universe is genuinely homogeneous, then when the radiation was last scattered, that temperature had to be the same everywhere. So the temperature of the radiation reaching us should be essentially the same in all directions.

However, because the Earth and the Sun have a random motion with respect to other galaxies, and therefore with respect to the mean rest frame of the galaxies, we

►Interestingly, *gravitational waves* are not scattered significantly by anything, and certainly do not care whether matter is ionized or neutral. Therefore, gravitational radiation *can* come to us from a much earlier time, and searching for a cosmological background of gravitational waves is one of the most important observations that gravitational wave detectors plan to make. If it exists, such a background would have originated from very much closer to the Big Bang than the cosmic microwave background, at a time when the Universe was less than 10^{-30} s old!

In this section: the microwave background defines a standard of rest. Astronomers have measured the Sun's velocity relative to it.

►Recall that a frame is an observer's coordinate system. So the rest frame of the Universe describes the preferred observer who is at rest in the Universe.

Investigation 24.2. Cosmic microwave background radiation in an expanding Universe

The cosmic microwave background radiation is a window into the early Universe. In this investigation, we will work out the two simple rules that govern how it changes when the Universe expands: (1) the wavelength of any photon remains proportional to the size of the Universe; and (2) the temperature of the photon gas is inversely proportional to this size. We will learn what the energy-density of the radiation is, and how to use it to measure our own speed relative to the cosmic rest frame.

We begin by asking how much energy the black-body radiation contains. We have seen in Chapter 10 that the total flux of energy emitted by the surface of a black body is given by the Stefan-Boltzmann law (Equation 10.11 on page 117):

$$F = \frac{2\pi^5 k^4}{15c^2 h^3} T^4. \quad (24.6)$$

This is the energy emitted per unit area per unit time. We want to use this to deduce the density of energy, the energy per unit volume.

Suppose that part of the surface is a flat plane with area A . If we look at the photons that are radiated by this area during a very short time-interval Δt , then they must all have come from the region just inside the black body and very near the surface. In particular, no emitted photon could have traveled further than $c\Delta t$ during the time Δt , so all the emitted photons originated inside a thin volume behind the surface of width $c\Delta t$. However, not all the photons that were in this volume at the beginning of the time-interval actually emerged from the surface. Half of them, in fact, had velocities directed *away* from the surface, back towards the interior, so they did not contribute to the flux. Of the rest, most were moving at an angle to the surface, so in the time Δt not all of them were able to reach the surface and contribute to the flux. It turns out (see Exercise 24.2.3) that the photons moving toward the surface have an *average* speed towards the surface of only $c/2$; the rest of their motion is parallel to the surface.

It follows from this that, of the photons that are in the thin volume, only one-quarter reach the surface and contribute to the flux. The energy that emerges is, therefore, only one-quarter of the energy in the volume. If the energy density of the radiation is denoted by ϵ_{bb} , then we have

$$\begin{aligned} \text{energy flux} \times \text{area of surface} \times \text{time-interval} &= \\ \tfrac{1}{4} \text{energy density} \times \text{volume of thin region}, \end{aligned}$$

which translates into the equation

$$FA\Delta t = \tfrac{1}{4}\epsilon_{bb}A(c\Delta t),$$

because the volume of the thin region is its surface area A times its depth $c\Delta t$. This can be solved for the energy density of black-body radiation:

$$\epsilon_{bb} = \frac{8\pi^5 k^4}{15c^3 h^3} T^4. \quad (24.7)$$

Exercise 24.2.1: Energy density of the cosmic microwave background

(a) Use Equation 24.7 to calculate the energy density of the cosmic microwave background, given its temperature of 2.7 K. (b) Show from this that the equivalent mass-density of the microwave background is $4.5 \times 10^{-31} \text{ kg m}^{-3}$.

Exercise 24.2.2: Motion through the cosmic background

According to measurements by COBE, the temperature of the cosmic microwave background has a maximum value on the sky that is 3.15 mK warmer than the average, and it has a minimum in a diametrically opposite direction that is 3.15 mK cooler than the average, after correcting for the motion of the satellite around the Earth and that of the Earth around the Sun. (The abbreviation mK stands for *millikelvin*.) Give an argument to show that the observed radiation should be black-body at a red- or blueshifted temperature. Then show that the speed of the Sun relative to the cosmic rest frame is $3.5 \times 10^5 \text{ m s}^{-1}$.

Exercise 24.2.3: Random photons

Devise a computer program, based on RANDOM, which allows you to calculate the mean speed toward the wall of the photons in the thin volume in our derivation of the energy density of a photon gas. Generate many random cases by choosing random directions for each photon (three random Cartesian coordinates) and calculating the speed toward the wall on the assumption that the total speed of the photon is c . Show that the mean speed toward the wall of those that move toward the wall is $c/2$.

Like the flux, this is proportional to T^4 .

Our next step is to consider a box containing black-body radiation of a given temperature T . The interior walls of the box are perfectly reflecting mirrors. Recall that, if this box has a tiny peep-hole, then the hole itself is a black body, since radiation falling onto the hole from outside will pass in, be reflected around the inside the box, and have negligible probability of re-emerging. The hole is a perfect absorber.

Now let us gradually expand the box until it is twice its original size in all dimensions. The hole in the box is still a black body, so the radiation inside still has a black-body spectrum. But what is its temperature?

As the box expands, the photons cannot keep their original wavelengths. A photon that encounters a moving wall will reflect in such a way that its energy is constant as measured by an experimenter at rest with respect to the wall, not by the experimenter at rest relative to the box. This results in a decrease in the energy of the photon at each reflection. This energy is lost in doing work to expand the box. The loss of energy produces a redshift of the wavelength of the light.

The redshift is proportional to the energy of the photon itself and to the speed of the wall. The redshift formula tells us that, for a single encounter with a wall, $(\Delta\lambda/\lambda)_{\text{single bounce}} \propto v/c$. In a very small time Δt , the photon will have many encounters, proportional to $c\Delta t/d$, where d is the size of the box. Its wavelength will increase according to the rule $(\Delta\lambda/\lambda)_{\text{during } \Delta t} \propto v\Delta t/d$. But the product $v\Delta t$ is the change in the size of the box during the expansion, Δd . So we have deduced the simple relation

$$\Delta\lambda/\lambda \propto \Delta d/d.$$

A more detailed calculation of the statistics of the photon velocities would show that the constant of proportionality here is just one. In other words, the fractional change in wavelength is the same as the fractional change in the size of the box. *If the box doubles in size, each photon gets stretched to twice its original wavelength.*

Since the spectrum remains black-body, and since the wavelength at which the spectrum reaches a maximum is inversely proportional to the black-body temperature, it follows that *the temperature of a black-body photon gas scales inversely with the size of the container. Therefore the energy density decreases as the inverse of the fourth power of the size of the container.* This scaling of the energy density is physically reasonable: the number of photons per unit volume is going down as the inverse cube of the size of the container, and the energy per photon has been redshifted by another factor of the size.

All of this happens to photons in the Universe. As the Universe expands, photons get redshifted. If the mean distance between galaxies expands by a factor of two then the cosmic background photons stretch to twice their original wavelength, and the energy density of the radiation decreases by a factor of 16.

would not expect to be exactly at rest with respect to the **surface of last scattering**: the distant sphere around us where the microwave photons finally became free. Instead, we would expect to be approaching that sphere in one direction and receding from it in the opposite direction. We would therefore expect to see a Doppler effect. The photons from one direction should be blueshifted, and therefore be hotter, and those from the opposite direction redshifted. This redshift preserves the form of the black-body spectrum but changes its temperature, as shown in Investigation 24.2.

This is indeed what is observed. The spectrum displayed in Figure 24.6 on page 354 is what results when the Doppler effects are removed. Scientists use the word **anisotropy** to describe the deviations from isotropy in the radiation. Measuring the anisotropy allows astronomers to calculate the Sun's velocity, as we describe in Exercise 24.2.2.

Our Sun's velocity through the Universe is measured by this means to be 350 km s^{-1} .

It is interesting to reflect that relativity began with the assumption that we could never measure our own velocity, not in an absolute sense. We discussed this in Chapter 15. Has relativity now led to the opposite, a measurement of our velocity in the Universe? There is no real contradiction here, but the issue is intriguing nevertheless. The velocity we measure from the cosmic microwave background is, like every other velocity in relativity, a velocity *relative* to something. In this case, it is relative to the mean rest frame of matter at the time the universe became transparent to photons. We only determined this velocity by making a measurement on something outside of ourselves: the radiation itself. Velocities are still relative, even in cosmology.

But what is intriguing about this is that there is only one Universe, so the Universe really does have a preferred rest frame. This is the frame in which the mass in the Universe is at rest, on average. And since the random velocities of galaxies are small compared with the expansion speed of the Universe over distances we can easily observe, this rest frame really is the best frame for describing the physics of the Universe. When we consider how nuclear physics determined the creation of elements after the Big Bang, or how galaxies formed from small density irregularities, or how the microwave background itself was formed: for all of these questions it is very helpful to do the physics in the preferred rest frame of the Universe.

Are there other variations in the temperature of the microwave background in different directions? There should be, since the Universe is irregular on small scales. But the variations are incredibly small.

After correcting for our motion, the temperature of the radiation in any given direction differs from that shown in Figure 24.6 on page 354 by an amount of order one part in 100 000.

This is an extraordinary degree of homogeneity. Once we take away the Doppler effect of our own motion, the radiation has the same temperature even when coming from regions that apparently could not have communicated with one another between the Big Bang and the time they emitted the radiation. The Copernican principle for observers at those remote locations seems very good indeed!

Big Crunch or Big Freeze: what happens next?

The engine at the heart of the Universe is gravity. Gravity is what makes the Big Bang slow down, gravity is what is making it speed up again, and gravity will decide its long-term future. If the Universe is indeed homogeneous and isotropic, then *gravity alone* will determine its future development. No other forces can act on

In this section: the future of the Universe depends on the way the dark energy behaves in time. It is likely to expand forever, but it might re-collapse.

▷ A good example of a local force that has no net result is the pressure *force*. Pressure forces act through non-uniformities, as we saw in Chapter 7. Therefore, they cannot directly accelerate the Universe's expansion, because there is no *net* pressure force on any part of the Universe.

▷ Interestingly, these three possible outcomes are similar to the three types of orbit around the Sun that we met in Chapter 4: the particle in an elliptical orbit moves outwards, and then comes back in; the hyperbolic orbit moves outwards forever, limiting to a non-zero speed; and the marginal case of the parabolic orbit, which moves outwards forever, but with a limiting speed of zero.

▷ An alternative to the Big Crunch and Big Freeze is the steady-state theory of Hoyle and Narlikar (Chapter 11). This gained support in the 1950s through 1970s, but the discovery of the cosmic microwave background and of evidence for the creation of elements in the Big Bang (see the next chapter) made the theory go out of fashion. The accelerating Universe is a further severe problem for this theory.

In this section: Newtonian theory is incomplete when describing cosmology.

masses in a homogeneous and isotropic Universe. If a force exists that pushes a bit of the Universe in a certain direction, then by isotropy there must be an opposing force of equal size pushing in the other direction.

Gravity is the only force that determines what happens to the Universe as a whole, and indeed what has happened in the past.

However, although local pressure *forces* are not important, pressure itself helps to create gravity by contributing to the *active gravitational mass* (Chapter 19). If the Universe is filled with matter that has negative pressure, then this can make the active gravitational mass negative, and then gravity will turn repulsive. This appears to be happening now, with a field of unknown origin contributing enough negative pressure to accelerate the Universe.

If we want to predict the future of the Universe, we have to have certain information about its present state. We need the density and pressure of the Universe, so we can calculate the density of the active gravitational mass and therefore the gravitational acceleration. And we need the present Hubble expansion speed of the Universe, from which we predict future expansion speeds using the acceleration. And finally, we need to know the physics of how the density and pressure change when the Universe expands.

Since we assume that the Universe behaves the same way everywhere, there are only two main types of futures, always expanding or eventually re-contracting; and there is the marginal class in between, where the Universe just manages to keep expanding, but with its expansion speed approaching zero. These are illustrated in Figure 24.8 on page 361. Can we say which evolution will be followed by our Universe?

Yes, we can, at least in principle. We will see below how we can use what we learned about escape speeds in Chapter 6 to calculate the Universe's escape speed, the expansion speed it needs now to go on expanding forever.

This analogy between orbits and cosmologies isn't perfect, of course; the Universe is not the same as a small satellite in the gravitational field of a planet. One difference is that the Universe provides its own gravity. Because of this, the bound universe does not cycle in and out the way an elliptical orbit does. Once it starts to re-collapse, the bound Universe shrinks toward infinite density, which is the time-reverse of its behavior at the Big Bang. Scientists have come to call this possible future the **Big Crunch**. Gravity is so strong at the Big Crunch that the Universe encounters a singularity: we cannot use the known laws of physics to predict what will happen after that. Many scientists hope that, if the Big Crunch happens, quantum gravity will get the Universe through the singularity and into another phase of expansion, but this is pure speculation at present.

The evidence today, however, is that the Big Crunch may not happen. Instead, the measured acceleration of the Universe suggests that the Universe is actually supplying its own *anti-gravity!* This makes it behave in a way that is unlike anything we saw with planetary or cometary motion. The acceleration, if it continues forever, will make the Universe bigger and bigger, colder and colder. The stars will eventually burn out, and the Universe will be cold and dark forever. This is the **Big Freeze**.

Cosmology according to Newton

If we want to know what might happen to our Universe in the future, we must study the laws of motion of an expanding universe, find the balance between expansion speed and gravity. Let us first see how this works using Newton's law of gravity.

The homogeneity of the Universe makes Newtonian cosmology almost as simple as studying the motion of a planet around the Sun. Homogeneity allows us to place ourselves at the “center” of the Universe and to calculate the forces on other galaxies that attract them towards us. All we want is to find the relative acceleration of the galaxy and ourselves; we do this by placing ourselves at the middle, so we have no acceleration, and calculating the acceleration of the galaxy due to its location. The key result that makes this calculation simple is as follows.

The net gravitational force on a galaxy that is a distance d away from us is produced only by the part of the Universe that is within a distance d of us. This force is the same as if all the mass within this distance d were concentrated at a point at our own location.

The rest of the Universe, further away from us than the galaxy, *has no net gravitational effect on it!*

The argument for this key result is basically the theorem of Newton that we proved using the computer in Chapter 4. As illustrated in Table 4.4 on page 35, the Newtonian gravitational force on an object inside a *spherical shell* is zero. Let us consider the part of the Universe that is further away from us than the galaxy we are considering.

If we can divide it up into a series of concentric spherical shells, we will have proved that it can exert no net gravitational force on the galaxy. If the Universe is truly homogeneous and infinite, then we can surely do this, so the result would follow. This idea is illustrated in Figure 24.7.

Unfortunately, this argument has a big problem. We appear to have to invoke the nature of the Universe arbitrarily far away: what happens if, in a part of the Universe so far away that we cannot yet see it, the mass distribution in the Universe actually comes to an end, and the edge is not spherically symmetric about our location? That would upset the argument, which required that we split the Universe up into exactly spherical shells centered on us. The ragged edge is far away but has a lot of mass, being distributed over a huge surface area, so it would have a measurable effect on gravity here. Is it acceptable that the evolution of the Universe in our neighborhood should depend on the detailed structure of the Universe outside the particle horizon? The problem here is clearly that we have pushed Newtonian gravity too far. Only in relativity, where regions very far away have not had time to affect us gravitationally yet, can this paradox be resolved.

Cosmology according to Einstein

General relativity is essential if we want to describe the Universe in the large. There are in fact at least two reasons for this. The first is the finite age of the Universe, which implies, as we have seen above, that there is only a finite portion of the Universe (inside our particle horizon) that can have influenced us. Gravitational influences in general relativity cannot travel faster than light. Indeed, gravitational waves travel at exactly the speed of light. This tells us that, in the argument above where we divided the Universe into spherical shells, we need only go out as far as the distance that light can have traveled since the Universe began. Anything further away has had no influence yet on gravity here. So the gravitational force on a relatively nearby galaxy cannot depend on whether the Universe is spherical in a

▷ In one respect cosmology is even simpler than planetary motion: bodies in the Solar System move around the Sun as well as toward and away from it, while in cosmology a homogeneous Universe only moves in and out.

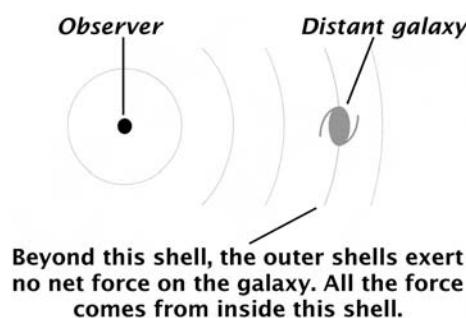


Figure 24.7. Dividing the Universe into concentric spherical shells, so that the gravitational force attracting a distant galaxy to us depends only on the mass closer to us than it.

▷ The problem of the influence of very distant regions is a serious one in Newtonian gravity, and is one of the reasons that cosmology did not become a serious study until Einstein provided a better theory of gravity.

In this section: Einstein’s theory allowed the first consistent physical theory of cosmology. But to calculate its evolution one can get away with equations from Newtonian gravity.

region we can't observe; it only depends on the Universe being homogeneous out to as far as we can see. The observations of the cosmic microwave background radiation reassure us that this is indeed the case.

The second reason for needing general relativity is that, if we consider galaxies so far away that their Hubble recession speed approaches the speed of light, then Newtonian gravity must fail to be valid. Let us remind ourselves of Hubble's law, that the speed of recession of a galaxy is proportional to its distance,

$$v = Hd, \quad (24.8)$$

where H is Hubble's constant. In this expression, if we can make d big enough, then we can make v bigger than the speed of light. Clearly the expression in Equation 24.8 must change if we go far enough away from our Galaxy, even in a homogeneous Universe. This expression can only be a *local* approximation to the recessions speeds we measure.

Extending the Hubble law to large distances might seem to be a big complication, but it is something we can postpone worrying about until after we have studied the basic dynamics of the Universe. The reason is as follows.

The expansion and contraction of a homogeneous Universe is a *local* phenomenon: as long as we can calculate the motion of *nearby* galaxies relative to us, then the homogeneity of the Universe guarantees that all other galaxies will behave the same way.

The key point, which we shall now argue, is that we can calculate the acceleration of nearby galaxies relative to us in Einstein's theory in the same *local* way, using concentric shells of matter, as we sketched in Newton's theory.

We have already noted in Chapter 21 that the gravitational field outside a spherical star in general relativity is identical to the field outside a black hole of the same mass. The inverse of this is also true: if we take a spherical distribution of mass in general relativity and cut a spherical hole in the middle, leaving the hole empty, then in the hole there will be no gravitational acceleration: spacetime will be perfectly flat, just as in special relativity.

So even in relativity, the gravitational acceleration of a galaxy relative to us depends only on the part of the Universe closer to us than it is. Now, if we consider only galaxies so near by that their recession speed is very much less than the speed of light, then we should be able to use Newtonian gravity to describe their motion. It follows that the dynamics of the Universe can be described by Newtonian gravity, provided that we use the correct relativistic source of gravity, which is the active gravitational mass.

Despite its logical inconsistencies, a spherical Newtonian cosmology is an accurate approximation to the relativistic cosmology if we use the correct relativistic form for the source of gravity.

Evolving the Universe

Here we shall see how to use Newtonian dynamics to find the dynamics of the relativistic Universe. First we will make a simplifying assumption and neglect all pressure, supposing that the Universe is dominated by the observed matter and the inferred dark matter. This is probably not accurate today, but it was a good approximation over much of the early evolution of the Universe, and especially at the time galaxies were formed. We shall therefore begin our study of the evolving Universe with these assumptions, and come back to the effects of pressure later. We call this the **matter-dominated** Universe.

▷ The inverse should be no great surprise: we saw in Chapter 19 that at least the dominant change in the way gravity is created in general relativity is that the Newtonian mass density is replaced by a combination of density and pressure, the active gravitational mass. Therefore, properties of the gravitational field that depend on symmetries, such as the way matter is distributed, should be the same in relativity as in Newtonian gravity.

In this section: we define and calculate key numbers, like the critical density and the density and deceleration parameters.

Let us use the symbol ρ to represent the mass-density of the Universe averaged over volumes that today are about 100 Mpc in size. This will be the same everywhere, by our assumption of homogeneity. A galaxy at a distance d from us lies on a sphere whose volume is $4\pi d^3/3$. The mass closer to us than the galaxy is then $M = 4\pi\rho d^3/3$. The part of the Universe further away contributes nothing to the net force, and the net force on the galaxy is the same as it would be if all this mass were concentrated at the center (our location).

Now, the key idea that keeps matter-dominated cosmology simple is that any particular galaxy at the distance d will *always* feel only the gravitational force from the same mass M , as long as pressure is negligible and the matter is non-relativistic. All the mass closer to us expands less rapidly, thus never going further away than that galaxy; all the mass further from us expands more rapidly, thus never coming closer to us than the galaxy; and none of the mass in the volume disappears. It follows that the escape speed of the galaxy is exactly the same as if the galaxy were escaping from a fixed point mass M . This speed is, as we saw in Equation 6.10 on page 56,

$$v_{\text{escape}} = \left(\frac{2GM}{d} \right)^{1/2}.$$

Substituting $4\pi\rho d^3/3$ for M in this equation, we find

$$v_{\text{escape}} = \left(\frac{8}{3}\pi G \rho \right)^{1/2} d. \quad (24.9)$$

This has the same form as the Hubble law, Equation 24.8, namely $v \propto d$. It tells us that if the proportionality factor H in the Hubble law exceeds the proportionality factor $(8/3)\pi G \rho^{1/2}$ in this equation, then *every* galaxy – regardless of d – will be going faster than the escape speed, and so the Universe will continue to expand forever. Put another way, if the Universe is matter-dominated with a Hubble constant whose value is H_0 , then there is a **critical density** ρ_c for which the Universe is just marginally bound. We find this value by setting the coefficient in Equation 24.9 equal to H_0 , squaring, and solving for ρ :

$$\rho_c = \frac{3H_0^2}{8\pi G}. \quad (24.10)$$

Taking $H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$, we find that the critical density today is about $\rho_c = 10^{-26} \text{ kg m}^{-3}$.

Let us compare this number with astronomers' estimates of the mean density of the Universe. The smallest possible value for ρ is the density we obtain if we spread out the luminous mass (not the unseen dark matter) of observed galaxies. This is estimated by astronomers to be no more than $5 \times 10^{-29} \text{ kg m}^{-3}$ if $H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$. This is 200 times smaller than the critical density. Cosmologists often prefer to couch their discussions of the mass density in terms of a dimensionless quantity, the ratio Ω of the true mass density to the critical density, called the **density parameter**:

$$\Omega = \frac{\rho}{\rho_c}. \quad (24.11)$$

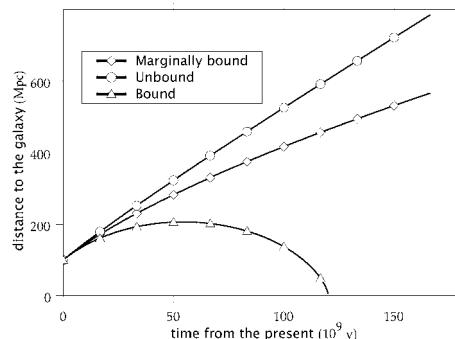


Figure 24.8. Three possible future developments of a simple matter-dominated universe that has a Hubble constant of $50 \text{ km s}^{-1} \text{ Mpc}^{-1}$. The re-collapsing universe model has a density that is twice the critical density. The other models shown have density equal to and half of the critical density, respectively. For definiteness, we plot the position at any time of a galaxy that starts at a distance of 100 Mpc from the origin.

►The fact that the mass in the volume is always the same is the key simplification of the matter-dominated assumption: when we consider a photon gas, whose energy redshifts away during the expansion, or a cosmological constant, whose energy density is constant with time, then the active gravitational mass inside the volume changes with time.

►The critical density defined this way is important even for cosmologies that are not matter-dominated. We will see in Chapter 26 that it determines the curvature of space.

►The symbol Ω is the capitalized version of the Greek letter omega, the last letter of the Greek alphabet.

The visible mass density contributes a fraction $\Omega_{\text{vis}} = 0.005$ of the critical density. By itself, the visible mass density could not stop the expansion of the Universe.

We saw in Chapter 14 that there is a lot of missing mass. We shall see in Chapter 25 that some of it – perhaps four times as much as the visible mass – is in the form of hydrogen and helium gas that has never formed stars. Much more than this is hidden dark matter, in a form that astronomers have not yet identified. The best estimates of the amount of dark matter on the cosmological scale give densities a factor of three lower than the critical density: $\Omega_M = 0.3$.

If the Universe at the present time were matter-dominated, then it would have less than the critical density, and it would expand forever. We calculate the actual deceleration of such a model in Investigation 24.3, where we address an important detail of principle: we show that the deceleration is also proportional to distance, so that the Hubble law (Equation 24.8 on page 360) remains true for all time. The result is the important equation:

$$a_{\text{cosmol}} = -\frac{4}{3}\pi G \rho d. \quad (24.12)$$

The fact that the inward acceleration increases in proportion to d implies that the Hubble law will hold for all time. Cosmologists usually write this equation in a slightly different way, defining the dimensionless **deceleration parameter** q by

$$q = \frac{4\pi G \rho}{3H^2}, \quad (24.13)$$

so that

$$a = -qH^2 d.$$

The value of the deceleration parameter today, q_0 , is related to the dimensionless density parameter Ω of the Universe, defined in Equation 24.11 on the preceding page, by

$$\Omega = 2q_0. \quad (24.14)$$

The expansion of the Universe at present appears to be accelerating, so that means it cannot be matter-dominated. To include pressure, one simply replaces ρ by the density of active gravitational mass, $\rho + 3p/c^2$, in Equation 24.12. In particular, the dark energy must be added to other mass densities when comparing with the critical density of the Universe. Since the dark energy behaves like a cosmological constant, astronomers denote its density relative to the critical density by the symbol Ω_Λ .

As Figure 24.4 on page 352 shows, the density associated with the dark energy is just enough, when added to the density of the dark matter, to give the critical density, within the observational errors. This result is borne out by measurements of the cosmic microwave background radiation as well, as illustrated in Figure 27.2 on page 403.

This is an unexpected result that many physicists want to see explained. As we will see in Chapter 27, it is predicted by the theory of inflation.

However, when pressure is important the evolution of the Universe will of course be different. In particular, the work done by the pressure as the Universe expands will affect the mass-energy density ρ . Moreover, the pressure itself can change. Normally, to decide whether a particular model Universe will expand forever or re-collapse requires a computer calculation.

Investigation 24.3. Cosmological gravity

A key question is, does the expansion of the Universe maintain the Hubble law? Hubble discovered the expansion in the first place by finding that the speed of recession of a galaxy is proportional to its distance,

$$v = Hd, \quad (24.15)$$

where H is Hubble's constant. If this describes the velocity of matter in the Universe now, then will the expansion of the Universe change it? After billions of years, will the expansion law look different, say with speed depending on the square of the distance? We don't expect it to, since the Hubble law is the only one that a homogeneous Universe can satisfy. But we need to check that the law of gravity does indeed maintain this. Otherwise, we have a logical inconsistency in our model of the Universe.

Now, the expansion of the Universe must be slowing down or speeding up, due to gravity, so Hubble's "constant" is generally not constant in time. If the Hubble law is preserved, then it follows that the acceleration (or deceleration) must also be proportional to the distance. Therefore, we expect to find, at least near to our Galaxy,

$$a = Kd, \quad (24.16)$$

where K is a different "constant", a number that is independent of position but can change with time. Because, as we have seen in the text, the acceleration depends only on the mass closer to us than d , we can calculate this in the same manner as we calculated the escape speed. The mass closer to us than the galaxy is $M = 4\pi\rho d^3/3$, and the acceleration it produces is $-GM/d^2$. This gives the cosmological acceleration:

$$a_{\text{cosmol}} = -\frac{GM}{d^2} = -\frac{4\pi G\rho d^3}{3d^2} = -\frac{4\pi G\rho}{3} d. \quad (24.17)$$

Exercise 24.3.1: The emptiness of the Universe

Since the luminous mass in galaxies is primarily in hydrogen, what would be the mean volume occupied by a single hydrogen atom if the mass in galaxies were smoothed out over the entire Universe? (Use the mean density of visible matter given in the text, $5 \times 10^{-29} \text{ kg m}^{-3}$.)

Exercise 24.3.2: Local accelerations

The nearest large galaxy to us is the Andromeda galaxy (also called M31), which is about 0.5 Mpc away and is falling towards our Galaxy, not receding from it. Take the mass of our Galaxy to be $10^{11} M_\odot$ and calculate the gravitational acceleration produced by our galaxy on the Andromeda galaxy, using the formula $a = -GM/r^2$. Calculate the cosmological acceleration given by Equation 24.17 at a distance of 0.5 Mpc, using the critical density ρ_c . Compare the two accelerations. Are motions within the local group of galaxies (those dominated by Andromeda and ourselves) strongly affected by the expansion of the Universe?

Exercise 24.3.3: Relation between Ω and q

Derive Equation 24.14 from Equations 24.11, 24.10, and 24.18.

Reassuringly, the acceleration increases in proportion to d , just as we expected: *our model of an expanding homogeneous universe governed by the known laws of gravity is self-consistent*. The constant of proportionality in Equation 24.16 is then

$$K = -4\pi G\rho/3.$$

Cosmologists do not usually deal with K directly. Instead, they define a dimensionless measure of the deceleration, called the *deceleration parameter*. Here is how it is defined.

We have already noted that the Hubble constant has dimensions of 1/time, and that $1/H$ is a measure of the age of the Universe. Now look at the proportionality constant K . Its dimensions are those of acceleration divided by distance, which works out to be $1/(\text{time})^2$. So the ratio K/H^2 is *dimensionless*. It is, to within a sign, what cosmologists call the deceleration parameter q :

$$q = -\frac{K}{H^2} = \frac{4\pi G\rho}{3H^2}. \quad (24.18)$$

We can thus write the cosmological acceleration in Equation 24.12 as

$$a = -qH^2 d.$$

The present values of all these "constants" are denoted by a subscript "0": H_0 , q_0 , ρ_0 , and p_0 .

Starting from Equation 24.17, it is possible to calculate the expected evolution of the Universe. We show how to do this with the help of a computer in Investigation 24.4 on the next page.

If pressure is not negligible, say because of radiation in the early Universe or because of a cosmological constant, then we just replace ρ by the relativistic $\rho + 3p/c^2$ everywhere in this calculation.

The cosmological scale-factor

Astronomers describe how a cosmological model expands by using the cosmological scale-factor. We introduced this for the rubber-band universe in Investigation 24.1 on page 348. The idea is a way to describe changes in the size of a cosmological model, even if the model is infinitely large. No matter how large the model is overall, if the distance between two typical galaxies doubles over some period of time, then the "size" of the cosmology has effectively doubled. These relative size changes are the important aspects of cosmological expansion, not the overall size of the Universe. The **cosmological scale-factor** tracks this relative expansion.

Consider two galaxies at some particular reference time, say at the present moment t_0 . Let their separation be d_0 . At another time t they are separated by a distance $d(t)$. We define the scale-factor R to be the ratio of these distances

$$R = \frac{d(t)}{d_0}. \quad (24.19)$$

In this section: the appropriate indicator of the expansion of the Universe is the change in relative distances between nearby points. This is directly measured by the cosmological redshift.

Investigation 24.4. Making the Universe grow

In this investigation, we shall see how to use a computer to evolve the equations for the matter-dominated Universe.

We start with an extension of the reasoning in Investigation 24.3 on the preceding page that led us to the critical mass density ρ_c . Every spherical shell surrounding us is affected only by the mass inside it. Since shells never cross, the mass inside it remains constant, so the fact that the portion of the Universe inside the shell is also expanding is irrelevant: it simply contributes a gravitational pull on the shell which is exactly like the pull on, say, a spacecraft launched from a planet of constant mass. Since the shell consists just of independent galaxies, all moving radially outwards, they all move on exactly the same trajectories that free particles would take if they were launched radially outwards from the same position with the same mass inside.

This makes it very easy to calculate the future development of the Universe: we just use our orbit program Orbit on the website, with initial numbers adapted to the present problem. For example, let us consider a galaxy that is 100 Mpc away from us at present, and let us take Hubble's constant to be $50 \text{ km s}^{-1} \text{ Mpc}^{-1}$. Then the galaxy

has an initial speed of 5000 km s^{-1} . We can orient our x -axis for this problem to be along the direction from us to the galaxy. Then it will have an initial x -speed of $5 \times 10^6 \text{ m s}^{-1}$ and an initial y -speed of 0. Its initial x -distance is 100 Mpc, or $3 \times 10^{24} \text{ m}$. Its initial y -distance is 0. If we take, as an example, a universe model which has half the critical density for this H_0 , then the mass closer to us than the galaxy is $4\pi(\rho_c/2)x^3/3 = 3 \times 10^{47} \text{ kg}$.

The program requires the variable K to be GM , which in this case is 1.9×10^{37} in si units. Since we will want to follow the Universe as it expands for something like the Hubble time of $6 \times 10^{17} \text{ s}$, we take a time-step that is much smaller than this, $5 \times 10^{14} \text{ s}$.

The curves in Figure 24.8 on page 361 show the results of this and two other simulations, one for a universe that has only half the critical density, and one for twice the density. The three graphs show how the position of the initial galaxy changes with time in each of three possible futures: re-collapse to a Big Crunch, expanding forever but slowing to zero speed, or expanding forever at a constant speed.

Clearly, R is a function of the time t , and of the reference time t_0 . But it is *not* a function of the galaxies we chose. The reason is Hubble's law: if the two galaxies had initially been twice as far apart, their expansion speed would have been twice as large, so the distance between them would have increased by twice as much, and the ratio of $d(t)$ to the original d_0 would turn out to be exactly the same.

So the scale-factor tells us how the Universe is expanding. We can express other physical quantities in terms of it. For example, the number of galaxies in our expanding Universe is not changing, but the distances between them are increasing in proportion to the scale-factor. The mean density ρ_M of matter in the Universe is decreasing in inverse proportion to the volume containing any given collection of galaxies. Since the volume is the product of three lengths, all of which are increasing in proportion to R , the volume is proportional to R^3 , and the density is inversely proportional to this:

$$\rho_M \propto 1/R^3.$$

The scale-factor is directly measurable in the cosmological redshift, and this is one of the most important relations in cosmology. The redshift is a Doppler shift, which we described in Figure 2.3 and Investigation 2.1 on page 15. If the light comes to us from a galaxy at a distance d , then the galaxy is receding with its Hubble speed $v = Hd$, and the redshift is $z = v/c = Hd/c$. The ratio of the wavelength the light has when we receive it, λ , to its original wavelength, λ_0 , is

$$\frac{\lambda}{\lambda_0} = 1 + z = 1 + \frac{Hd}{c}. \quad (24.20)$$

Now consider what has happened to the Universe during the time the photon was moving from its source galaxy to us. The photon took a time $\tau = d/c$ to travel to us at the speed of light. In this time, its source galaxy moved from its original distance d to a further distance $d + v\tau$. Therefore, the scale of the Universe has increased by the factor

$$R = \frac{d + v\tau}{d} = \frac{d + (Hd)(d/c)}{d} = 1 + \frac{Hd}{c}.$$

This is identical to the wavelength ratio in Equation 24.20 above. We have therefore verified, by a different method, and for photons that are not part of the cosmic microwave background, the same remarkable and simple result as we had for the expanding photon gas in Investigation 24.2 on page 356:

►We won't see the galaxy at its new location until the photons it emits "now" have had time to reach us, of course.

the wavelength of a photon that moves freely through the Universe increases in direct proportion to the cosmological scale-factor R .

Expressed as an equation, this is

$$\lambda \propto R. \quad (24.21)$$

This allows us to look back in time and make conclusions about what the early Universe was like. For example, if we examine a quasar whose redshift is four, then the scale-factor of the Universe was only one-fifth (because $1 + z = 5$) of what it is today. The average distances between galaxies were only 20% of what they are today, and the temperature of the cosmic microwave background radiation was five times higher than today, or around 13 K.

What is the cosmological expansion: does space itself expand?

Because of the cosmological redshift, photons behave as if they were being stretched between the galaxies: their wavelengths increase exactly as the distances between galaxies increase. It might be tempting to conclude from this that the cosmological expansion stretches space itself, just as our rubber-band universe was built out of a stretching material. But this could be quite misleading from a physical point of view.

In particular, one must not think that *space itself* is everywhere enlarging, as if extra “points” were somehow being created among the old ones all the time, and everything was getting bigger in proportion to the Universe: wavelengths, sizes of atoms, sizes of people. If the sizes of atoms and the spacings between atoms in molecules were getting bigger, then the expansion of the Universe would be unmeasurable. If our rulers were enlarging at the same rate as the wavelengths of photons from distant quasars, then we would not notice the redshift, since the incoming light would occupy the same fraction of the standard meter as it did when it left the quasar billions of years ago. The expansion of the Universe is an observable fact precisely *because* ordinary matter does not expand with it.

It is simpler from a physical point of view to think of the expanding Universe as a simple collection of particles (called galaxies) that are rushing away from one another. The redshift of light is a Doppler shift caused by the motion of the source galaxy away from us.

Notice how this looks from the point of view of a photon. Let us take its source galaxy as the standard of rest, the (arbitrary) center of the Universe. As it moves away from the source, it passes other galaxies. They are all moving away, but not as fast as the photon, which moves at the speed of light. The further the photon travels through the Universe, the faster is the speed of the galaxies it passes, since the faster galaxies have traveled further from the center since the Big Bang. Suppose that it happens to be detected by an astronomer in one of these galaxies. Then, the longer it has traveled, the faster will be the speed of the detecting galaxy relative to the source galaxy, and the bigger will be the detected redshift. The redshift increases with time, but this increase has nothing to do with a metaphysical stretching of space: it is simply the way the Doppler shift works in a homogeneous expanding Universe.

Why does ordinary matter not expand with the Universe? After all, each proton and electron starts out from the Big Bang participating in the cosmological expansion. The answer is that the particles would continue to expand if they remained free particles, influenced only by the smooth cosmological gravitational field. But they don't remain free.

The other forces of Nature, such as electromagnetism, disturb the cosmological expansion in small regions, binding individual particles to one

In this section: the expansion of the Universe is not a mystical expansion of space, but rather the expansion of distances between ordinary objects.

another, wiping out the initial relative velocity between them. Once the “memory” of the initial expansion is lost, atoms are governed by forces in their neighborhood, not by cosmological gravity.

►Physicists who use supercomputers to simulate the formation of galaxy clusters – as a way of testing the cold dark matter hypothesis – use Newtonian gravity in their simulations, because it is so much easier to use than general relativity, and it is a perfectly adequate approximation. See Figure 25.3 on page 380 for the results of one such simulation.

This remark even applies to irregularities in gravity. Galaxies form from the expanding gas of the Universe because of some random irregularity in the density of the expanding gas, which causes a local increase in gravity that slows the expansion of the nearby gas. Eventually, if the initial irregularity is big enough, the local self-gravity is strong enough to reverse the cosmological expansion in the gas, and the gas becomes a gravitationally bound object, perhaps a galaxy or a cluster of galaxies. After that it can be described perfectly adequately by Newtonian gravity, ignoring the rest of the Universe. The *average* motion of the cluster of galaxies can't be wiped out by the forces between the particles in the cluster, so it still participates in the cosmological expansion, as a whole cluster, relative to other clusters and galaxies. But within the cluster, the expansion has been forgotten.

The age of the Universe

In this section: the computed age of the Universe depends on the value of the acceleration today. The age is probably between 12 and 13 billion years.

Let us again look backwards in time, to the Big Bang. We saw that the Hubble constant gives us an upper bound on the age of the Universe provided it was matter-dominated over most of its past. With a more realistic model for the Universe, one can get a better estimate of its age.

The early history of the Universe was very complicated, but this period was short: the Universe became matter-dominated after about 400 000 y. In recent times, however, things have become complicated again: the acceleration of the expansion has dominated for about the last half of the age of the Universe. This makes the estimate of the age very sensitive to the assumed value of the acceleration. The best estimate today is that it is between 12 and 13 billion years old. This is consistent with the estimated ages of all known stars and clusters.

But is this enough time to produce the Universe as we see it? Can stars and galaxies form in this amount of time, do galaxies have enough time to clump into rich clusters? This is not an easy question to answer. In fact, without dark matter, the answer would be no: not nearly enough time. We will see in the next chapter how the clumpiness of dark matter helps accelerate the formation of structure in the evolving Universe.

The Big Bang: the seed from which we grew

We have now pushed our model of the history of our Universe back just about as far as we could hope to go: the Universe had a beginning, and that beginning was the source of all that happened afterwards – all matter, all stars, all galaxies, even life itself. Big Bang cosmology has placed modern physics in the remarkable position of being able in principle to trace back to the beginning every aspect of the world we live in, to say “This is where X came from, and this is how Y started”.

Physicists have grasped this opportunity with enthusiasm. The study of what is often called **physical cosmology** – the evolution of the matter in the Universe after the Big Bang – is one of the most active and exciting branches of astrophysics today. Helped by powerful computers, physicists can now explain how the elements hydrogen and helium were made, where the cosmic background radiation came from, and (at least in outline) how galaxies and clusters of galaxies might have formed. Within these galaxies, we have already seen in Chapter 12 how stars arose and turned hydrogen and helium into all the heavier elements, and in Chapter 7 we speculated on how a tiny portion of these heavier elements became the planet Earth and produced the conditions that allowed an even tinier portion to become living things.

We can, still very imperfectly, trace our own origins as humans right back to the beginning of the Universe.

This chapter is about *physical cosmology*: what happened in great arena of the Universe from the beginning to now.

Physical cosmology: everything but the first nanosecond

Let us imagine running the movie of the Universe’s expansion backwards, so we observe it getting denser and denser. How far can we go and still claim to understand what is going on? The answer is startling: physicists can go back with confidence to within 10^{-10} s of the Big Bang. And they can even make some shrewd guesses about what happened as early as 10^{-35} s.

The reason for this remarkable success is the homogeneity of the Universe. Going back in time is just like compressing the matter in the Universe into smaller and smaller volumes. This makes the matter hotter, and the particles in the matter become more energetic. As long as the typical particle energy is less than the limits of our understanding of particle physics from man-made accelerators (about 1 TeV), we can be confident of our description of what happens at these energies and densities. By carefully solving the Einstein equations for the evolving Universe, generalizing what we did in Investigation 24.4 on page 364, scientists find that the average particle energy fell to about 1 TeV at about 10^{10} s after the Big Bang. This epoch marks a watershed in physicists’ models of the Universe: before this time, they speculate; after this time they speak with fair confidence. We will look in Chapter 27 at the most interesting current speculations about the earliest phase of the Universe. In this chapter we deal with what is known with some confidence.

In this chapter: we study physical cosmology: how physics worked in the expanding Universe. This includes the formation of the elements hydrogen and helium, the role of dark matter, and the formation of galaxies and clusters. Physicists have achieved a remarkable understanding of the Universe after its first nanosecond.

▷ The background image on this page is a plot of the spatial distribution of galaxies in a thin wedge of space centered on our position, from the CfA Redshift Survey. Measured in 1985, this distribution gave astronomers their first indication that galaxies were grouped into chains as well as clumps. The human-like pattern (horizontal in this view, with the legs to the left) became a celebrity in its own right. Data from M Geller and J Huchra, image copyright South African Observatory (sao).

In this section: physicists are fairly confident they understand the Universe after its first nanosecond!

▷ We met the electron volt, eV, as a unit of energy in Chapter 8. High-energy physicists usually describe energies as multiples of this unit: keV (10^3 eV), MeV (10^6 eV), GeV (10^9 eV), and TeV (10^{12} eV).

In this section: the earliest stage that is well-understood is the quark soup: when the Universe was still too hot and dense to allow protons and neutrons to exist individually.

▷ This is nothing more than the kinetic temperature: express the energy E of 1 TeV in ordinary units (joules) and set it equal to $3kT/2$ to find T . You will get about 10^{16} K.

The expansion of the quark soup and its radiation

What was the Universe like at 10^{-10} s? If the Universe is highly homogeneous today, it must have been more so at these early times, since all the structure we see today (galaxies, and so on) developed at a later time as the Universe cooled down. We will describe the way structure developed later in this chapter. But little of it was present when the Universe was a hot ball of gas with a typical particle energy of 1 TeV and a corresponding temperature of 10^{16} K.

Ordinary matter as we know it could not exist at these temperatures. If there were any ordinary protons around, then their random collisions would break them up into their constituents, which physicists call quarks. Quarks are the oddest particles in physics: in groups of three they make protons or neutrons, and in groups of two they make π -mesons and other lighter particles. Yet one never sees them alone: single quarks cannot be peeled off from particles in accelerators. In the early Universe, the particles were packed so closely together that quarks were never alone. Instead, they blended together in a sea that physicists call the **quark soup**.

Besides quarks, there were many other particles in the early Universe. Whatever particles now constitute the dark matter were already there, but their density irregularities, which would be important for galaxy formation later on, were not significant at this time. The dark matter particles were neutral and had stopped interacting with the quarks or the photons by this time. They were already just a provider of a gravitational background.

And there were photons. With energies typical of the thermal energy, they had enough energy to form new quarks in reactions where two photons collide and two quarks emerge. By mechanisms like this and the reverse, the numbers of quarks and photons were maintained in a steady balance.

When particles like quarks or protons are produced by photons that collide with one another, they emerge with equal numbers of particles and anti-particles. The anti-particle of any particle has the opposite sign of the charge. So if electrons are produced, one is a normal electron and the other is a positron, or positively charged electron. If a proton is produced, an **anti-proton** (with a negative charge) is also produced. The anti-particle of a photon is just another photon. So in this way, no net charge is produced: the two photons initially have zero charge, and the two particles that emerge have zero total charge.

When an anti-particle and a particle of the same type collide, the result is often to produce a pair of photons, which is the time-reverse of the reaction described in the previous paragraph. Thus, a proton and anti-proton will annihilate each other to produce two photons. Similarly, a positron and an electron annihilate to two photons. Physicists refer to positrons, anti-protons, anti-neutrons, anti-quarks, and so on collectively as **anti-matter**.

The laws of physics prefer matter over anti-matter

In this section: if the laws of physics were perfectly symmetrical between matter and anti-matter, all matter would have been annihilated and we would not be here. We owe our existence to a small preference in the laws of physics for matter over anti-matter.

As the Universe expanded, the mean distance between quarks grew until they began to get too isolated. When this happened, they started to clump into twos and threes, forming ordinary protons, neutrons, π -mesons, and other particles. The corresponding anti-quarks also clumped to form anti-protons, anti-neutrons, anti- π -mesons, and so on.

As the Universe cooled further, the photons, whose gas stays at the same temperature as the particles because they collide frequently, no longer have enough energy to create proton-anti-proton pairs when they collide. At this point, there are still lots of collisions where protons and anti-protons annihilate to form photons, but the photons get redshifted by the expansion of the Universe and, by the time they meet other photons, no longer have enough energy to create protons and anti-protons

again. This also applies to neutrons and anti-neutrons. The number of particles decreases steeply at this point because annihilations are dominant over creations.

In principle, all the protons should have annihilated against all the anti-protons. But in fact, it is obvious that they did not: we are all made of protons that survived this era. It is natural, then, to expect that some anti-protons also survived, but this apparently did not happen. The two were so well mixed that we should see anti-protons everywhere, and we don't. Instead, it appears that there were simply more protons than anti-protons, by a small amount. This can only reflect a fundamental asymmetry in the laws of physics, a preference for one kind of matter against its opposite.

As the Universe expanded and cooled, the same thing happened later for the electrons: when the temperature was too small to create electron–positron pairs, then the electrons and positrons annihilated. The asymmetry at this point was exactly the same: the same laws of physics allowed the same fraction of excess electrons to survive as for protons.

We can learn how slight the excess of protons was by counting photons today. The microwave background radiation has on average 10^9 times more photons in any region of space than there are protons and neutrons. This number has not changed much since the separation of photons and electrons and the subsequent annihilations took place. The number of electrons has not changed, and the number of photons has changed only by a factor of two or so by the processes we describe below. It follows that the excess of protons/electrons over anti-protons/positrons in the very early Universe was about 10^{-9} . This is a small clue to the nature of laws of physics that physicists do not yet understand. We will return to this point in Chapter 27.

We owe our existence to this slight asymmetry in the laws of physics.

If the laws were perfectly symmetrical between matter and anti-matter, then all the protons would have been annihilated in the early Universe, and there would have been nothing left to build stars, planets, and people from. The Universe would instead have been filled with pure radiation, cooling as it expanded.

It is interesting to reflect that we are formed from the waste that resulted from a slight imperfection in the laws of physics!

The Universe becomes ordinary

The annihilation of protons and neutrons stopped when the thermal energy kT fell to about the rest-mass-energy of a proton $m_p c^2$. It is easy to calculate that this temperature is about 10^{13} K. Since this temperature is smaller by a factor of 100 than the quark-soup temperature we quoted above, it follows that the mean photon energy had gone down by a similar amount, and therefore that the Universe had expanded by the same factor of 100. By Equation 25.1 on the next page, the time since the Big Bang had therefore increased by a factor of 10^4 , to 10^6 s.

The electrons annihilated at a temperature of about 6×10^9 K, when the Universe was a further factor of 1600 larger, and a factor of 3×10^6 older. It was now 3 s old. This is the epoch at which ordinary matter appears.

After the first microsecond, nuclear matter was already the normal material of which the nuclei of all elements is made. After the first three seconds, all the remaining exotic particles had disappeared, and the Universe was made of familiar stuff.

Notice how much physics takes place in times that seem short to us: everything really exotic is finished in the first three seconds! When one deals with the early

In this section: the excess protons and neutrons eventually dominated the composition of the early Universe, accompanied by electrons and neutrinos. Most other particles had gone away after the first few seconds.

Universe, one quickly begins to think, not in terms of time-intervals, but in terms of time *ratios*.

The time-interval between the end of the anti-proton era at 10^{-6} s and its beginning at 10^{-10} s should not be thought of as $0.9999 \mu\text{s}$; rather, it should be thought of as a ratio: time since the Big Bang has increased by a factor of 10^4 , and the scale-factor by a factor of 100.

This way of thinking about cosmological time is shown in Figure 25.1 on page 375. This displays time logarithmically, which insures that time-intervals that have the same ratio are separated by the same distance along the line.

After the annihilation of the anti-protons, the total energy density of the Universe was dominated by the photon gas. The remaining particles had little total energy, and by colliding frequently with photons they kept the same temperature as the photons. We say that the Universe was radiation-dominated.

One can show (see Investigation 25.1) that a radiation-dominated cosmological model, at least soon after the Big Bang, expands in such a way that its scale-factor R is proportional to the square-root of the time t since the Big Bang:

$$R(t) \propto t^{1/2}. \quad (25.1)$$

What happened to the energy in the photon gas as the Universe expanded? The number of photons remained approximately constant, so the number of photons per unit volume was determined only by the volume of the clump, which increased as R^3 . Just as we saw for galaxies in the previous chapter, the number of photons per unit volume was proportional to R^{-3} .

But we also saw there that the wavelength of a freely moving photon in an expanding cosmology increases with R . Therefore its frequency ν is inversely proportional to R , and its energy $h\nu$ is inversely proportional to R . So the *energy density* of the photon gas (energy per photon times number of photons per unit volume) was proportional to R^{-4} . Expressed as an equation, this is:

$$\text{energy density of a photon gas} \propto R^{-4}. \quad (25.2)$$

This is a deep result, and it applies to any period of time when photons are either dominant over matter or move freely without scattering from matter. In particular, it holds for the cosmic microwave background radiation today.

Besides the photons there were also neutrinos in the early Universe. Observations of neutrinos from the Sun now strongly suggest that the neutrino has a non-zero rest-mass, but they don't tell us yet what it is. But the rest-mass energy equivalent will be less than a few electron volts, very small compared to the average neutrino total energy in the early quark soup. So the neutrinos at this epoch would all have been traveling essentially at the speed of light. Moreover, the density of the soup was so large that the neutrinos scattered frequently, keeping them in equilibrium with the quarks and hence with the photons. So the neutrinos at this time formed a gas with a temperature and energy density similar to the photons, and there were likewise billions of neutrinos for every quark.

Making helium: first steps toward life

In this section: nuclear reactions stopped as the Universe expanded and cooled. Their main product was helium. Most of the helium in existence today was formed in the Big Bang.

We, as human beings, have a real interest in what happened in the early Universe: if things had been very different, and the right conditions for life had not emerged, then we would not be here. One of the indisputable essentials for life is the existence of heavy elements, such as carbon and oxygen. These are made in stars from the basic building blocks of helium nuclei. But the helium itself is not made in stars

Investigation 25.1. Exact solutions for marginally bound model universes

Here we determine how the scale-factor R of the Universe depends on time for different kinds of assumptions about the composition of the Universe. Our derivation will use rather simple ideas, but we will arrive at accurate answers.

We begin with the fundamental equation for the acceleration (or deceleration) of a galaxy at a distance d away from us in a matter-dominated Universe, Equation 24.12 on page 362, which we reprint here:

$$a_{\text{cosmol}} = -\frac{4}{3}\pi G \rho d.$$

In such a Universe, the density ρ is proportional to $1/R^3$, because as the Universe expands the particles simply get spread out over a larger volume. The location d of any particular galaxy is also proportional to R , so the right-hand side of this equation is proportional to R^{-2} .

The left-hand side is harder to work out. First, as we have just noted, the galaxy's distance from us is proportional to R . Therefore its velocity away from us is the rate of change of R . Without knowing how R depends on time yet (that is the goal of this calculation!), let us make the crude approximation that the speed of the galaxy is proportional to R/t . This approximation is exact if the speed is constant and R increases in direct proportion to t . If R is proportional to, say, t^2 , then the approximation suggests that the speed is proportional to t . This is in fact correct, as we saw in the discussion of gravity on the Earth in Investigation 1.2 on page 4. Going another step further, we can guess that the acceleration of the galaxy will be proportional to R/t^2 , again exactly as in Investigation 1.2.

The result of these guesses is that the acceleration equation implies

$$R/t^2 \propto 1/R^2.$$

We can multiply by R^2 and t^2 and solve:

$$R^3 \propto t^2, \quad \Rightarrow \quad R \propto t^{2/3}.$$

This turns out to be exactly correct: the scale-factor of a matter-dominated Universe is proportional to the $-2/3$ power of the time. You might worry that the method is crude, and indeed it is. In a moment we will show how we would be able to tell whether the method was giving a wrong answer. In this case, we are getting the right answer for a *particular* solution to the equation.

Our solution is not the most general one possible. We have explored the general solutions with a computer, and the results are in Figure 24.8 on page 361. In fact, since according to our present solution, R increases unboundedly with time, this solution cannot represent the bound Universe. Also, since the rate of change of R , which is proportional to R/t , goes to zero as t gets larger and larger, it also does not represent the unbound Universe. We are left with just one possibility: we have found the law governing the marginally bound solution in Figure 24.8 on page 361. Notice, however, that in this figure all three solutions behave very like one another in the early Universe. Therefore, we have also found a good approximation to any matter-dominated cosmology in its early phase.

If we want to examine other kinds of Universe evolutions, we must replace ρ by $\rho + 3p/c^2$ in the acceleration law:

$$a_{\text{cosmol}} = -\left(\frac{4\pi G}{3}(\rho + 3p/c^2)\right) d.$$

For example, the radiation-dominated Universe is a Universe filled with a photon gas. The pressure of such a gas is proportional to its density, in fact $p = \rho c^2/3$. So the right-hand side of this equation is proportional to ρ , just as in the matter-dominated case. However, as the Universe expands, the energy density of the photons decreases more rapidly. As we can infer from Equation 24.21 on page 365, the energy per photon decreases as $1/R$, while the number of photons per unit volume decreases in the same way as the number of

particles in a matter-dominated Universe, by $1/R^3$. This means that the right-hand side of this equation is proportional to $1/R^4$ for the radiation-dominated Universe. In Exercise 25.1.1 below, we show that this leads to the law $R \propto t^2$. Again this is the solution representing the marginally bound case, but it is also a good approximation to all the solutions in the very early Universe. This is particularly relevant, since the very early Universe (after any period of inflation) was radiation-dominated.

Let us try this reasoning on the case where the Universe is dominated by a cosmological constant, for which $p_\Lambda = -\rho_\Lambda c^2$. Then the right-hand side of the acceleration equation is proportional to ρ_Λ , which is a *constant*. The minus sign on the right-hand side has been cancelled by the minus sign that comes from $\rho_\Lambda + 3p_\Lambda/c^2 = -2\rho_\Lambda$, so that the acceleration is *positive*. This is how the cosmological constant produces an accelerating Universe. If we follow the same method for finding how R depends on time, we get

$$R/t^2 \propto R,$$

since the only dependence on R on the right-hand side is now the factor d . If we try to solve this for R , we get into trouble: R divides out, leaving us with the non-sensical result $t^2 \propto 1$.

So our method fails for this type of Universe. The approximation that the acceleration is proportional to R/t^2 is not consistent with the acceleration equation. We shall see how to find the right answer in a moment. But let us point out in passing that the fact that the method did not fail for the matter-dominated and radiation-dominated cases tells us that we can believe the answers that we got, however crude our initial guess was. Essentially, our initial guess did not have to be right, but if the equation gives a consistent result after making the guess, then this retrospectively confirms that the guess was correct.

So how do we proceed when the guess is wrong? We make another guess, of course. To see what new guess might be reasonable, let us first ask a simpler question. What if our equation was not for acceleration but for velocity? What if it said that the rate of change of R was proportional to R ? This is a familiar situation in lots of physical problems. It happens in radioactivity, for example: the number of nuclei that decay in a given time is proportional to the number that are there. It happens to populations of rabbits, as well, if they are not limited by availability of water or food or by predators or disease: the number of new rabbits in a given year is proportional to the overall number of rabbits.

Such problems are known as *exponential* problems, and they have solutions in which the number of things (nuclei, rabbits) is proportional to e^{kt} , where k is a constant that is not determined by these arguments. We met the exponential function when studying the black-body law in Chapter 10.

Now let us go back to our problem with acceleration. Do we still have exponential behavior? The answer is yes. Again, let us assume that the scale-factor of the Universe increases exponentially in time. Then we have just seen that its rate of change is proportional to itself and therefore increases exponentially in time. This in turn means that the acceleration, which is the rate of change of the expansion speed, is also proportional to the scale-factor. This is exactly what the acceleration equation gives for the cosmological constant. Therefore, we have guessed a consistent solution this time, and it is the right one: *the Universe expands exponentially when dominated by the cosmological constant*.

The theory of inflation postulates that the Universe went through a phase of exponential expansion at a very early time. Recent observations of supernovae suggest that our Universe has recently entered this kind of phase again. If the laws of physics give us a cosmological constant that really is constant for all time (see Chapter 27), then our Universe will expand exponentially. Exponential expansion is very rapid: the bigger it gets, the faster it goes!

Exercise 25.1.1: Radiation-dominated universe

Find the dependence of the scale-factor on time for the radiation-dominated Universe. The analysis is similar to our derivation for the matter-dominated Universe above. The only difference is that, as explained above, the factor $\rho + 3p/c^2$ is proportional to R^4 . Show from this that $R \propto t^2$.

►The fact that we *are* here tells us something already about the early Universe: it had to make some amount of helium, for example. Reasoning of this kind is related to the Anthropic Principle, which we mentioned in Chapter 11 and to which we will return in Chapter 27.

in great quantities – the Big Bang supplied this basic ingredient. Here is how it happened.

We have seen that after the first microsecond, the matter in the Universe consisted mainly of protons, neutrons, neutrinos, photons, and the electron–positron gas. The positrons had little effect on nuclear reactions, so we ignore them here. Now, neutrons on their own are unstable particles: as we observed in Chapter 12, the neutron is slightly more massive than a proton and an electron, so it can decay into them and release some energy, which is carried away by a neutrino. But in the early Universe, electrons had so much energy that when they collided with protons there was more than enough to turn the pair into a neutron. So after 1 μs , neutrons and protons were more or less in equilibrium with one another, and with the electrons and neutrinos as well.

This happy situation could not last forever, because the expansion of the Universe was relentlessly cooling things off. Eventually the density became low enough so that the neutrinos stopped scattering from the electrons and protons. This occurred at about 10^{-2} s. From that time until now, the cosmological background of neutrinos has been cooling off. Its temperature now is about 2 K, but since neutrinos are so much harder to detect than photons, it has never been directly observed.

The photons were still scattering off protons and electrons, of course, and insuring that all the particles stayed at the same temperature. All kinds of other collisions were also happening. Protons and neutrons would collide to form a deuterium nucleus, for example, but soon afterwards a photon would collide with the nucleus and break it apart. So some of the protons and neutrons were to be found in light nuclei at any time, but not many.

However, once the expansion had cooled off the photons sufficiently, they no longer had enough energy to break up the light nuclei that were constantly forming briefly by the collisions of protons and neutrons. The energy required to split up a deuterium nucleus is about 2 MeV. The equivalent temperature to this energy ($E = kT$) is 2×10^{10} K. Therefore, once the temperature of the photon gas had fallen below about 2×10^{10} K, there were not many photons around that could break up deuterium, so the random collisions of protons and neutrons quickly began to build up a density of deuterium. The deuterium nuclei occasionally suffered further collisions, and this led to the formation of helium nuclei, primarily ^4He , which consists of two protons and two neutrons.

Other light elements were formed at this time, up to ^7Li . But the expansion of the Universe reduced the density of these elements as rapidly as they could form, so nuclear reactions did not go beyond lithium in any quantity. The neutrons that were left free at this time (not inside deuterium or helium nuclei) subsequently decayed.

All of these nuclear reactions took place while the positrons were still present in large numbers. So at the end of the positron era at 3 s, the Universe contained protons, electrons, a decreasing population of free neutrons, and some nuclei. This was the gas out of which the first stars were made.

Does it correspond to reality?

In this section: we review the observational evidence supporting this picture of the early Big Bang.

Our story of the Big Bang so far has been based mainly on two sets of facts: astronomical observations of the expansion and homogeneity of the Universe, and our knowledge of laboratory nuclear and high-energy physics. But the calculations of helium formation make detailed predictions about the amounts of these elements that we should see around us, and these lead to independent checks on the theory. If the Big Bang predictions were seriously wrong, we would have to throw out at least

some parts of the theory. In fact, as we shall see, the predictions are so good that they led astronomers to new results in high-energy particle physics that were later verified by accelerator experiments. The result has been an enormous strengthening of our confidence that the Big Bang model provides a good description of the history of the Universe, at least back to 1 s after the Big Bang.

Detailed calculations of the synthesis of the light elements at around $t = 1\text{ s}$ are done by extensive computer calculations, but they also require one to make assumptions about a few numbers that astronomers do not have direct evidence for. One of these numbers is the density of protons and neutrons (nucleons) at the time of helium formation. Since most matter today seems to be dark, we can't simply trace the present density backwards in time. Instead, scientists calculate the predictions of the amounts of these elements produced by the Big Bang for a number of different values of the density of nucleons at 1 s, and compare the predictions with observational evidence about the amount of each element that the first generation of stars contained. These observations are done by taking spectra of stars and gas and measuring the strength of the lines in the spectra that are characteristic of the elements we are looking for.

This comparison has to be done carefully, since it is not easy to identify which stars and gas clouds (if any) belong to the first generation and which ones formed later from material "contaminated" by the waste products of the first generation. The problem is made easier by a key fact: deuterium and lithium are not created in the nuclear reactions that take place inside stars or when stars explode. So if astronomers observe deuterium and lithium in stars today, their abundances set a *lower bound* on the amount that was produced by the Big Bang.

When observations of lithium and deuterium are combined with observations of helium in the oldest stars known, the result is a fairly tight constraint on the amount of nucleons that were available to make these elements at about 1 s after the Big Bang. It tells us that *today* the cosmological density of nucleons is

$$\rho_{\text{nucleon}} = 0.14 \pm 0.03 \text{ nucleons per cubic meter.} \quad (25.3)$$

So if we were to spread the nucleons in the Universe smoothly out over its volume, there would be about one particle in every seven cubic meters! An average-sized room of 30 m^3 would contain just 4 hydrogen atoms. Multiplying by the mass of a proton, the nucleon mass density of the present Universe is $2 \times 10^{-28}\text{ kg m}^{-3}$. This is only 2% of the critical density needed to turn the Universe around, if the present Hubble constant is $70\text{ km s}^{-1}\text{ Mpc}^{-1}$. In the language of the previous chapter,

$$\Omega_{\text{nucleon}} = 0.02.$$

Small as this is, it is much larger than the density of the luminous matter that astronomers observe directly, which is around $5 \times 10^{-29}\text{ kg m}^{-3}$ for this value of the Hubble constant. So the production of He in the Big Bang tells us that perhaps 80% of the nucleons in the Universe are dark. This is remarkably consistent with the numbers that we get independently from studies of galaxies and clusters of galaxies, as described in Chapter 14. This is a further strong argument that the Big Bang is a good description of the early Universe.

Three and only three neutrinos: a triumph for Big Bang physics

These helium-formation studies have produced an even more remarkable test of their validity: astrophysicists were able to determine from them how many different kinds of neutrinos there are in nature. In order to produce the observed amounts of light elements, one does not just need exactly the right amounts of protons and neutrons to be available for the nuclear reactions. One also has to have the right

In this section: cosmologists predicted from the helium abundance that there could be only three kinds of neutrinos. Experimental particle physicists subsequently proved it.

expansion speed of the Universe at that time. If the Universe expands too fast, it produces less helium because the nuclear reactions turn off too quickly. In that case, there would be more deuterium left over. So the balance among the different elements today also tells us the expansion rate at the time of nucleosynthesis.

Now, one might think that the expansion rate today (Hubble constant) would tell us what the expansion rate was at the time of helium formation, but it is not so simple. The expansion rate since then has been slowing down because of the gravity of the Universe. To know the expansion rate at 1 s, we also have to know the self-gravity of the Universe. That depends on how much mass-energy there was at different times.

Today of course there is great uncertainty about the amount of mass, but fortunately that uncertainty does *not* affect the expansion rate at 1 s as much as one might expect. As we have seen earlier, the self-gravity of the Universe at 1 s was dominated by the *radiation* in it, not the particles. If the radiation was just photons, then by observing the microwave background today we could tell what the total self-gravity was at 1 s, and we could deduce the expansion rate of the Universe at that time. But we have left out the neutrinos in the Universe: they also behaved as a radiation gas at the time of helium formation, so their density contributes to our conclusions about the expansion rate then.

Now, at 1 s a neutrino gas would have been in equilibrium with the photon gas, so it would have had the same temperature and energy density. This would be true for each type of neutrino. Particle physicists have direct evidence for three kinds of neutrinos: electron neutrinos produced when an electron and a proton combine to form a neutron; mu neutrinos produced by the decay of a mu meson, or muon; and tau neutrinos produced by the decay of a tau meson. Each sort of neutrino would have formed a gas, so the density of the Universe at 1 s would have been at least four times the density of the photon gas itself.

But suppose there were a fourth kind of neutrino that particle physics experiments have not yet turned up. Particle physics theories allow this, and in fact some particle physicists have preferred theories with more than three kinds of neutrinos. Then the density of the Universe at 1 s would have been larger again, its self-gravity correspondingly larger, and the expansion speed it would have required at that time in order to reach the Hubble rate today would also have been larger. This would have quenched the helium production faster. The amount of helium decreases if there are more families of neutrinos.

Even taking into account the uncertainties in the present total mass density of the Universe, astrophysicists found that the only way to fit the observed amounts of all the light elements today was in a Universe that had exactly *three* kinds of neutrinos and has a present density of nucleons given by Equation 25.3 on the preceding page.

More recent particle physics experiments have also shown that there are only three kinds of neutrinos. By observing the decay rate of the Z^0 particle, which can only decay into particles accompanied by neutrinos, and whose decay rate is therefore proportional to the number of different possible neutrino species available to it, particle physicists at CERN found that there were only three decay modes.

This confirmation of the conclusion that had already been drawn on the basis of Big Bang cosmology was a real triumph for the Big Bang model, and it has led to a great deal of collaboration since then between astrophysicists and high-energy physicists to see what further light cosmology can shed on the behavior of particles at very high energies. The Big Bang is one of the few places that energies above

Physical History of the Universe

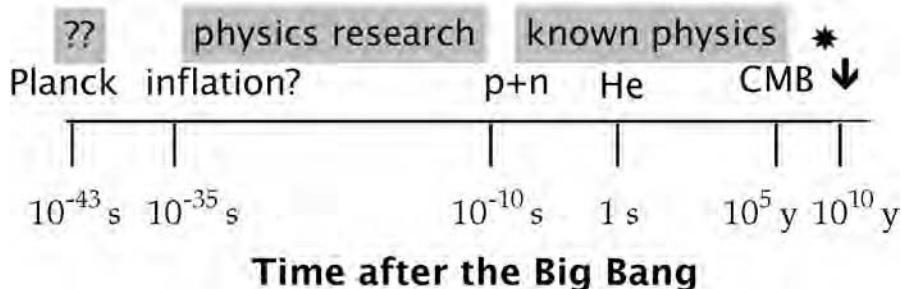


Figure 25.1. A time-line encapsulating the main features of the evolution of the Universe. Starting at the Planck time of 10^{-43} s, physicists know very little about the laws of physics governing the Universe until about 10^{-35} s, where matter began to dominate over anti-matter in the expanding ball of energy. The physics of this process is a principal area for research in high-energy physics today, so the period up to about 10^{-10} s is one where physicists have some understanding of how the Universe behaved. At this time, electrons and neutrinos began to behave differently from one another, and we enter the realm of well-understood physics. Soon after this the protons and neutrons condensed from the quark soup, and by 1 s the anti-protons and anti-neutrons had annihilated. Nuclear reactions formed helium and deuterium. Just before 1 s, neutrinos stopped scattering and became a free gas, and shortly thereafter the positrons annihilated. The cosmic microwave background (CMB) became a free photon gas at about 3×10^5 y, and the mass condensations that led to galaxy clusters began forming soon afterwards. The most distant quasars and galaxies that we see were shining and forming their earliest stars at the time *, and we live today near ↓.

those reached by our present accelerators have ever been seen, so this interest is natural. Now that physicists have confidence in the basic correctness of the Big Bang model, they can use it to illuminate other branches of physics. We shall see more of how this works in the Chapter 27.

From nuclei to atoms: the Universe goes transparent

Once the nuclear reactions had finished, there followed a long period in which the Universe simply cooled off as it expanded further. Initially, of course, all the matter was fully ionized: the nuclei had formed, but the energy of the particles was too large to allow the electrons to be bound to the nuclei. As we noted in Chapter 10, it takes only about 13 eV of energy to remove the electron from a hydrogen atom. As long as the average energy of the particles in the Universe is bigger than this, any electron that does get trapped by a nucleus will be knocked off it almost immediately. So the electrons will remain free.

Now, photons only scatter off charges. Photons are basically little packets of oscillating electric and magnetic fields, and these fields are affected by charged particles but not by neutral particles. Neutral atoms can scatter photons, but only if the photon can get close enough to "see" the individual charges within them. So they are not nearly as effective as free electrons and nuclei. The early Universe was a black body – a perfect absorber of photons – for the same reason that we found stars to be black bodies in Chapter 10. The ionized plasma traps the photons. Since the photons were still the dominant source of gravity, their temperature decreased inversely with the scale-factor R , and the matter temperature followed suit. The neutrinos also cooled off in the same way, so that even though they had stopped exchanging energy with other forms of matter, they continued to have the same temperature.

But eventually the temperature had to fall to the point where the typical energy was about 13 eV. This is when we would expect atoms to start forming and staying bound. But in fact, this was not the temperature at which most of the matter be-

In this section: once the nuclear reactions had stopped, the next big event was the formation of atoms. This required the gas to cool to below the ionization temperature of hydrogen. After this time the gas of the Universe is largely neutral, so light propagates without scattering. Our observations of the cosmic microwave background therefore go back to this time, at which radiation and matter decoupled.

came neutral. Recall our discussion in Chapter 10 of the ionization of hydrogen in the outer layers of the Sun, and especially why the surface temperature of the Sun is lower than the temperature where the typical particle energy would be 13 eV. The situation is the same in the early Universe, only more extreme, because there is a huge imbalance between the number of photons and the number of nucleons. With 1 billion photons for each nucleus, if only a small fraction of the photons have energies above 13 eV they can keep the matter ionized. In any gas, there is a random spread of energies. Not until the temperature of the photon gas fell to about 0.6 eV were there too few photons to keep the nuclei ionized. This is the epoch of decoupling.

The temperature ratio from the helium formation epoch to this time is the same as the ratio of the characteristic energies, $(2 \text{ MeV})/(0.6 \text{ eV}) = 3 \times 10^6$. Because of the relation between the temperature of the photon gas and the scale-factor, this is the ratio by which the Universe expanded between these two times. Since the elapsed time is proportional to the square of this ratio in a radiation-dominated Universe, the time has increased by a factor of 9×10^{12} . This puts us at $t = 9 \times 10^{12} \text{ s}$, or about 3×10^5 years.

Three hundred thousand years after the Big Bang, the Universe finally became neutral. After this time, it became largely transparent. The radiation that is now the microwave background was formed at this time and has been cooling off ever since.

The temperature at decoupling was the equivalent of 0.6 eV, which is about 4000 K. The microwave background temperature today is about 2.7 K. Therefore the photon gas has redshifted by a factor of about 1500 since decoupling. This is then the factor by which the Universe has expanded since then. The density of matter in the Universe has changed by the cube of this factor, about 8×10^9 . If we take the density today that is indicated by the helium-formation arguments, then this density at decoupling was about $2 \times 10^{-18} \text{ kg m}^{-3}$. This is already a very low density compared to everyday densities on the Earth, which are 10^{21} times larger. Forming the Earth therefore required a great concentration of matter at later times.

Coincidentally, the end of the plasma era is accompanied by another change: the transition from radiation-dominated to matter-dominated evolution. As the Universe expands, it is inevitable that this transition will take place. The energy of each photon decreases as the Universe expands, while matter particles have a reservoir of energy that does not go away: their rest-mass energy. So eventually, no matter how many photons there are, their total mass-energy will drop below that of the matter. The rest-mass-energy of a nucleon is about 10^9 eV, and there are about 10^9 photons per nucleon. The cross-over, therefore, occurs when the photons have an average energy of 1 eV. By coincidence, this happens at about the same time as decoupling.

After decoupling, the self-gravity of the Universe is dominated by matter. The background radiation of photons and neutrinos follows the expansion of the Universe but does not dominate it.

The evolution of structure

In this section: how the dark matter began the formation of galaxies.

Once matter becomes the dominant source of self-gravity, the details of the expansion and deceleration change somewhat, but the general trend was the same. The time since decoupling has been spent developing structure: this is the era during which clusters of galaxies, galaxies, and stars formed. This was a very complex

physical process, which astrophysicists are now coming to understand, aided by simulations performed on the world's largest supercomputers. Despite this complexity, there are some key aspects of the problem that we can consider in this book and draw conclusions about.

The first issue for us is to try to understand how, in a homogeneous Universe, irregularities like galaxies could have formed at all. Fundamentally, galaxies, stars, and planets all owe their existence to the basic fact about gravity that we mentioned on the first page of this book: it is universally attractive. In a smooth, expanding gas, this leads to instabilities. Any small irregularity might grow through its self-gravity. What actually happens is that in a region of higher-than-average density, the cosmological expansion slows down; the region continues to expand and get less dense, but the density *contrast* with the average density of the Universe gets bigger. If the region containing the density contrast is large enough, or the contrast is big enough, the region can actually reverse its expansion and re-collapse. Then we have a potential galaxy cluster. We will study one way in which this might have happened below.

This leads directly to our second question: where did the density irregularities come from? Of course, nothing is perfectly homogeneous. The positions of atoms are random, and that inevitably leads to clumping. But, as we show in Investigation 25.2 on the next page, the random clumping of atoms is too small ever to explain how vast numbers of them could have come together into clusters of galaxies. Something had to provide larger-scale density irregularities. Fortunately, we know that the Universe also contains dark matter whose form is undetermined. It is natural to look to the dark matter to provide these irregularities, rather than try to invent yet another mechanism.

As we saw in Chapter 14, the dark matter could come in several different forms, all of which make different predictions about the nature of galaxy clustering. The leading dark matter candidate at present is a sea of heavy, electrically neutral particles: this has come to be called **cold dark matter** (CDM). We will shortly see why this is attractive. Another candidate is cosmic defects, left over from exotic particle physics processes in the early Universe. This includes cosmic strings and cosmic textures. We will discuss these in Chapter 27, but here we only need to note that they may have concentrated considerable energy into small regions, providing mechanisms to start the collapse of ordinary matter.

The third question we can answer concerns the time at which galaxies might have begun to form. It seems certain that ordinary matter – nucleons and electrons – did not participate in any structure formation until after decoupling. The reason is that the ionized matter was very closely tied to the photon gas, so any clumping would have had to involve the photons too. But photons do not stay in one place, so they don't clump for long. Once they diffused away, the particle clumping would similarly die out. So before decoupling, any clumping could only have involved dark, uncharged matter. This is in fact the great advantage of invoking dark matter to start galaxy formation; it can get started much earlier than ordinary matter can. Galaxies only began to form after decoupling, but they formed by falling onto clumps of dark matter that had formed long before.

This brings us to the fourth issue, which is one of the main uncertainties about galaxy formation: why did the dark matter clump, and how clumpy did it get before decoupling? We have shown in Investigation 25.2 on the following page that random fluctuations in particle positions cannot account for any sensible degree of clumping. The dark matter accounts for more mass than the ordinary matter, so is this problem not even harder for dark matter?

Investigation 25.2. Can random clumping of particles lead to galaxy formation?

The first reason one might offer as an explanation of the density inhomogeneities that led to galaxy formation is pure randomness. One would expect, even in a homogeneous Big Bang, that particles would have small irregularities in their locations, in the same way that the molecules of the atmosphere we breathe are not perfectly uniformly distributed even though the atmosphere when averaged over many particles is uniform.

However, random irregularities in the density due to such effects get small when there are a lot of particles. Essentially, for every particle that moves closer to another, there is likely to be another that moves further away. It is a bit like the random walk that we studied in Chapter 8, and it has the same statistics: if a given small region of space would have, on average, N particles, then random fluctuations in particle positions will change it typically by a number of order $N^{1/2}$. The density contrast produced by a fluctuation, which is the ratio of the density fluctuation to the average density, is of order $N^{1/2}/N = N^{-1/2}$, which gets smaller as N gets larger.

Now, the number of particles that one needs to make even a star, let alone a galaxy or a cluster of galaxies, is very large. It is the ratio of the mass of the Sun to the mass of a proton, something like 10^{57} , so random fluctuations in particle positions will provide such a col-

lection of particles with a typical over-density of no more than one part in about 10^{23} . In order to form a star, the slight excess gravity of this fluctuation has to amplify the density contrast to something of order one.

This is simply much too small an initial contrast for this to have happened by now. Cosmologically speaking, stars have not had a lot of time to form. Gravitational collapse could not have started until after the Universe cooled enough for its highly ionized matter to have become neutral. Before that, the electrons scattered off the photons easily, which kept them too hot to collapse. Only when they re-combined with protons, and the photons no longer had enough energy to scatter from them, did matter have a chance to start collapsing. This was the time of the formation of the cosmic microwave background. Since then, the Universe has expanded by only a factor of 1000 or so. This would not have been enough time for random inhomogeneities to grow by a factor of 10^{23} .

Since the 1950s scientists have recognized that they had no obvious explanation for the initial fluctuations that led to galaxy formation. One reason that the theory of inflation is so attractive to many cosmologists is that it offers a natural explanation for bigger density fluctuations.

Exercise 25.2.1: Random clumping

Experiment with random clumping using a tossed coin as your random-number generator. Use three successive tosses to generate a number between 0 and 7, using its binary representation. That is, if the coin comes up heads assign a 1 to a digit, and if tails a 0. With three tosses you get three digits, say 010, and that is the number 2. (The digits abc represent the number $4a + 2b + c$.) Record each such number you get. Generate a large set of them, say 80. Each number should come up on average ten times, but some will come up more often and some less, at random. The excess over the average should be, according to the argument above, $10^{1/2} \approx 3$. You should expect some numbers to come up at least 13 times, and others only 7. You might expect one bin to have twice as large a fluctuation, i.e. to reach 16 or 4. Now go on and do twice as many, 160 numbers. (You need 480 coin tosses to do this!) Then the average will be 20 and the expected fluctuation $20^{1/2} \approx 4.5$. Although the fluctuation is larger in this case, it is a smaller fraction of the average, so that the distribution of numbers among the bins is actually smoother. If you have the stamina, go to 320 numbers. Verify that the typical fluctuation is of order six.

The clumping mechanism for dark matter depends on what the dark matter consists of. For the most popular CDM model, random particle positions are no help, because there will be a similarly huge number of these particles as of baryons. Instead, physicists tie the CDM model to the idea of inflation, which we will study in Chapter 27. Inflation has the curious property that it amplifies density irregularities by a huge factor. Even a small density fluctuation from quantum uncertainties before the era of inflation can be amplified into a significant irregularity in the distribution of dark matter particles after inflation. We shall have to wait until Chapter 27 to see how this works.

If the dark matter is in cosmic strings, then the strings themselves are large-scale objects, so it might be thought that they would form points of attraction for ordinary matter easily. However, the situation is a little more subtle. We will see in Chapter 27 that cosmic strings have zero active gravitational mass: their huge density is exactly cancelled by the equally huge negative pressure, which acts in only one direction, so that the active gravitational mass is $\rho + p/c^2 = 0$. They do not curve time at all, so they do not directly form places where matter clumps. That is, a static string sitting in the middle of a cloud of gas does not pull the gas towards itself. Instead, strings curve space, and this can only be felt by matter that moves transversely across the string. If the string moves through space, as it would be expected to do, then matter flowing around it on one side is brought by this deflection into collision with matter flowing around it from the other side. These collisions could cause the over-densities that lead to the formation of galaxies. Cosmic strings could form galaxies in the wake they leave as they move through ordinary matter.

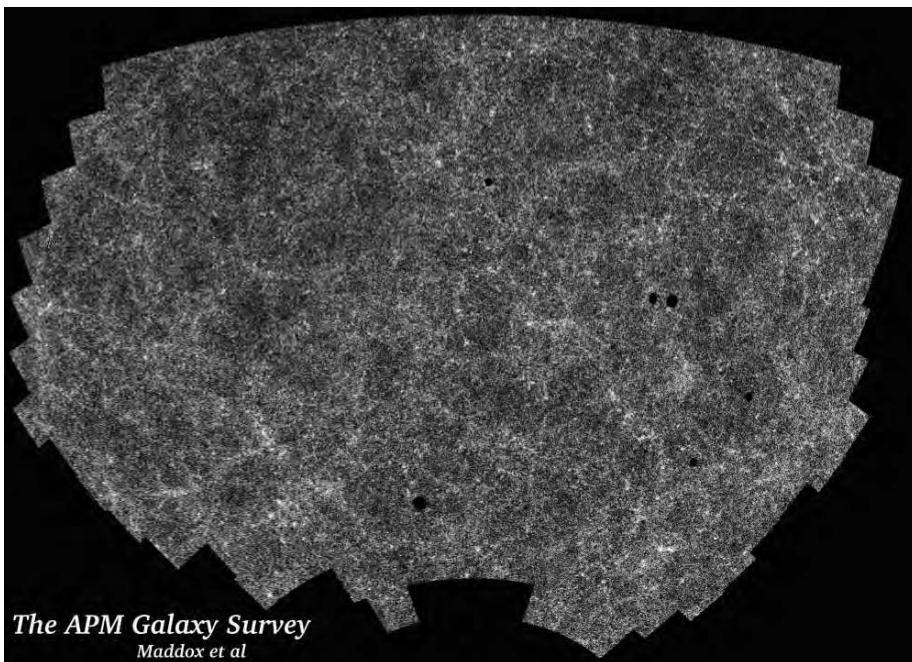


Figure 25.2. The galaxy distribution of the APM survey of a large region of the sky. The brightness of each point represents the number of galaxies there. The background is a smooth gray because of very distant galaxies, which are seen at an early time, so they do not clump very much. The brighter knots and filaments are made of nearby galaxies, which we see at a later stage of galaxy clustering. In fact these knots are superclusters, clusters of clusters, and the filaments between them are chains of clusters. Image courtesy of Steve Maddox, Will Sutherland, George Efstathiou and Jon Loveday, Astrophysics Department, Oxford University.

Ghosts of the dark matter

Theories of how dark matter forms galaxies can be tested against observations in at least two ways, even without directly detecting dark matter particles in the laboratory. The first way is to look for special characteristics in the way galaxies clump. The second way is to look for traces of the dark matter's density irregularities in the cosmic microwave background. Both these tests currently strongly favor CDM.

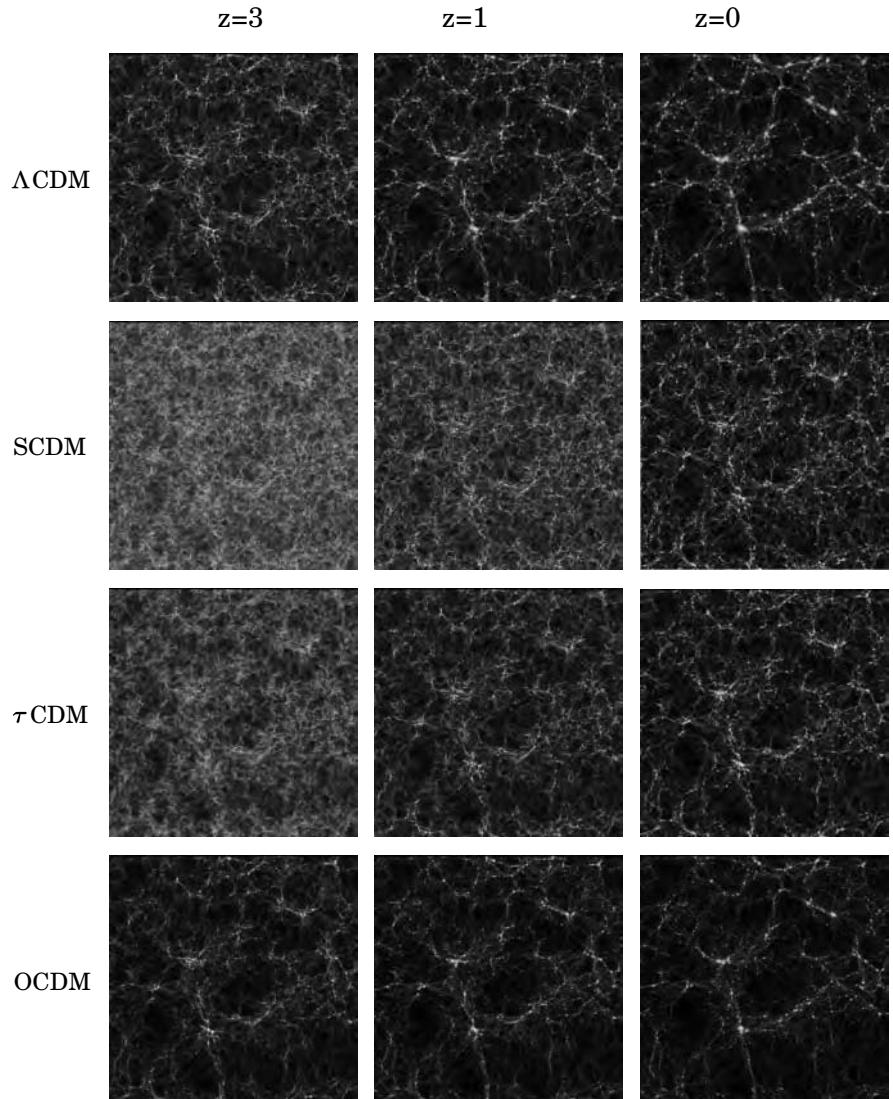
Studies of large-scale statistics of galaxy positions have shown that, not only do galaxy clusters clump into larger superclusters, but the superclusters tend to be connected by linear filaments which are also locations of large numbers of galaxy clusters. That is, galaxy clusters form a kind of web-like distribution in space, with superclusters at the knots in the strings of the web. This is illustrated in Figure 25.2, which is a computer plot of the positions of millions of observed galaxies across a large portion of the sky.

Statistics of clustering that are inferred from this kind of survey can be compared with computer simulations of what kind of clumping one might expect in the CDM model. Figure 25.3 on the following page shows the results of four such simulations, viewed at three different times. Notice how similar the structures at the present time in the first row of the figure look to the structures in white in Figure 25.2. This is obtained with a CDM model in which the dark matter has a density equal to 30% of the critical density (as defined in the previous chapter), and a cosmological constant with 70% of the critical density, as is suggested by the microwave background studies we look at next. This simulation fits the observations better than ones with other assumptions.

The clumping of dark matter also has a small but measurable effect on the microwave background radiation. The temperature of the radiation is lower toward a strong clump than elsewhere. This is a gravitational redshift effect, but with a subtlety. If light falls into a gravitational field and then leaves it, we have seen that its path is deflected (gravitational lensing) but we did not mention any redshift. That is because there is no net redshift: the energy of the photon is conserved if the gravita-

In this section: observations today of the spatial distribution of galaxies and of the irregularities in the cosmic microwave background radiation give clues to the nature, density, and distribution of dark matter.

Figure 25.3. Comparison of four simulations of galaxy clustering, with different assumptions about the cold dark matter. Each simulation is shown at three different times (redshifts) so that one can see how clustering gets stronger with time. Redshift $z = 3$ represents a time when the Universe was only one-quarter of its present size. Redshift $z = 1$ is when the Universe was one-half of its current size. Redshift $z = 0$ is the present time. Each image represents the same galaxies, so they are not shown to scale. In principle, one should reduce the images in the first column by a factor of four and those in the second by a factor of two, in order to see the expansion as well as the clustering. The top row is the preferred model, which produces clustering most closely like that in Figure 25.2 on the previous page and similar surveys. Its CDM has 30% of the critical mass, and it uses a cosmological constant with a further 70% of the critical mass density. The other models have different parameters. The second row, for example, is a model with no cosmological constant and a density of CDM equal to the critical density. The bottom row has the same CDM as the first but no cosmological constant. Figures made for the VIRGO consortium, by Joerg Colberg, published in Jenkins et al., 1998 *Astrophysical Journal*, 499, 20–40.



tional lens is static. However, in the cosmological case, the gravitational field of the clump is getting stronger during the time that the photon passes through it, so the photon's energy is not conserved. (Recall our discussion of energy conservation and time-dependent gravitational fields in Chapter 6 in association with the slingshot mechanism.) The gravitational field is stronger when the photon leaves, so it loses energy and is redshifted.

This means that very precise measurements of temperature irregularities in the cosmic microwave background can give information on the density irregularities that were forming at the time of decoupling and later. The first such measurements were made in the early 1990s by the COBE satellite. We showed the overall spectrum measured by COBE in the last chapter in Figure 24.6 on page 354. COBE also measured the temperature fluctuations and showed that they have a size consistent with what

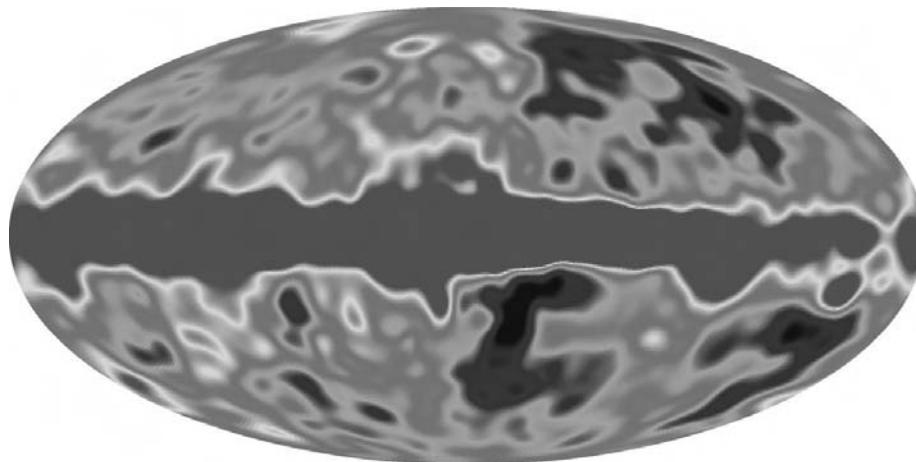


Figure 25.4. This map shows the irregularities in the temperature of the cosmic microwave background in different directions on the sky, as measured by the COBE satellite. The mean temperature has been subtracted; the fluctuations shown here are only of order 10^{-5} of the overall temperature. Courtesy COBE team and NASA/GSFC.

is needed to explain galaxy formation. The results, shown in Figure 25.4, give a striking visual representation of the CDM fluctuations before galaxies formed. Even better pictures are expected soon from the MAP satellite, which was launched in 2001. Later in the decade we can expect the launch of the most sophisticated cosmic microwave background satellite of all, called Planck.

Since COBE, more accurate measurements have been made by instruments that have been flown on high-altitude balloons. These experiments, called Boomerang and Maxima, examine fluctuations on much smaller angular scales, which are even more closely related to the galaxy count maps such as Figure 25.2 on page 379. These observations strongly favor the distribution of galaxies that would be produced by the CDM model with inflation, and are beginning to exclude cosmic strings as the main cause of galaxy formation. These experiments also have given independent information about the cosmological constant. We will see that evidence in Figure 27.2 on page 403.

What is the dark matter?

While galaxy clustering and the cosmic microwave background observations independently point toward the inflation with CDM model, they do not themselves tell us much about the CDM particle that is involved. All that is required to match observations is a neutral particle that does not easily scatter from baryons and whose mass is sufficiently large that the expansion of the Universe has cooled the random velocities of the particles enough to clump. As far as these results are concerned, the particles could be either as-yet undiscovered elementary particles (which we called WIMPs in Chapter 14) or, say, massive black holes.

The reason that scientists favor the former is that the element-formation studies that we described above require that the number of baryons should be a small fraction of the critical density, while the dark matter is a much larger fraction. So if the dark matter is in black holes, then either the black holes formed from some non-baryonic particles (in which case we still need unknown particles), or the black holes formed from baryons before the epoch of helium formation. Since the Universe was hot and smooth at that point, forming black holes would either require some exotic quantum process at the end of inflation, or it might require a very radically different model of the very early Universe in which there was a cold Big Bang. Neither of these options is simple, so physicists will continue to assume that CDM is an elementary particle until it is found in laboratory experiments or until experiments show somehow that these particles do not exist.

In this section: most astrophysicists favor the idea that the dark matter is made of uncharged elementary particles that do not feel the nuclear force. Other models are possible, and intensive searches are underway to identify the particles.

>Theoretical physicists would not be surprised to find WIMP particles. Modern theories suggest a large collection of so-called supersymmetric particles, many of which could have the right properties for CDM.

One of the components of the dark matter seems to be neutrinos. The accumulating evidence from studies of solar neutrinos is that neutrinos have a small mass, which is not zero but is smaller than 1 eV. Given, as we saw above, that neutrinos should be as abundant as microwave background photons, this is a significant amount of matter. Neutrinos could form up to 5–10% of the dark matter.

But neutrinos cannot be responsible for galaxy formation, because they were not “cold” at the time of decoupling. At that point, the neutrino temperature was of the same order as the photon temperature, an energy equivalent as we have seen above of a few electron volts. With rest-masses much smaller than this, neutrinos would have been moving at close to the speed of light at decoupling, and they could not have formed the stable, tight clumps needed to start galaxy formation. So neutrinos constitute **hot dark matter**: interesting, but not sufficient to complete our picture of the composition of the Universe.

The hunt for this elusive component of the Universe is one of the most interesting experimental activities today. Deep in underground laboratories, a number of groups of scientists are monitoring very sensitive equipment for evidence of unexpected particle events that cannot be explained by known particles. The labs are underground to screen out cosmic rays, which would otherwise create such a large background of events that the desired ones would be hard to identify. Some groups are looking for new kinds of nuclear reactions; others simply look for the tiny heat and sound waves generated by a collision between a dark matter particle and the material of the detector. Experiments are getting more sensitive all the time and have begun to put constraints on dark matter models. A direct detection of a dark matter particle, and a solution of the mystery of how they fit into our modern picture of particle physics, may be only a few years away.

Einstein's Universe: the geometry of cosmology

In the last two chapters we have made a lot of progress in exploring the future and past of the Universe, basically just by using local Newtonian gravity. We argued that the dynamics of an expanding, homogeneous and isotropic cosmology can be calculated from Newtonian gravity, at least if the pressure in the Universe is negligible, because all we need to look at is the local Universe, the part nearest us. The assumption that the Universe is homogeneous guarantees that the rest of the Universe will behave the same as our local region.

But this line of reasoning has its limitations. Even if we calculate the dynamics of the Universe this way, we don't learn what the distant parts of the Universe will *look like* in our telescopes. The curvature of space, which is not part of a Newtonian discussion, will affect the paths of photons as they move through the Universe. Moreover, if we want to ask deeper questions about the Universe, such as those we pose in the next chapter, then we should know something more about its larger-scale structure. For this, we must turn to full general relativity. Only general relativity can provide a consistent picture over the vast scales we shall need to explore, out to where the cosmological speed of recession approaches the speed of light. So it is now time to learn about Einstein's description of cosmology.

Cosmology could be complicated ...

As we have seen, Einstein's theory has the simplifying property that only matter within our past light-cone – matter that can send signals to us – can have influenced the evolution of the Universe we observe. This is logically much more satisfying than Newtonian gravity, where matter everywhere affects us with its gravity instantly. In fact, scientists did not study cosmology seriously before Einstein: the logical difficulty of applying Newtonian gravity to an infinite Universe, coupled with the fact that astronomers before the twentieth century had no idea how large the Universe was, left scientists with little to work with. When Einstein's theory showed how to treat gravity in a causal way and provided consistent cosmological models, scientists began to explore the subject.

The basically Newtonian view of cosmology we developed in the last two chapters was still based on relativity: we had to use the two facts that (1) only matter in our past light-cone affects our gravitational field, and (2) general relativity allows us to ignore the gravity due to spherical mass distributions further away from us than the galaxy whose motion we are computing. For homogeneous universe models, we were then able to ignore most of relativity and study the dynamics with essentially Newtonian equations. We will develop below the relativistic counterparts of these model universes, and we will see that in many situations they are remarkably similar.

But relativity is richer than Newtonian gravity. There are model universes that are not describable in Newtonian terms. Here is an example.

Imagine a homogeneous universe in which the expansion is different in different directions. For example, imagine that the Universe were expanding at twice the rate

In this chapter: we explore the three different geometries that a homogeneous and isotropic cosmology can assume. We see how to construct two-dimensional versions of these, which shows us why there are only three possibilities. We see how astronomical observations can measure this geometry directly.

►The drawing under the text on this page illustrates how complicated three-dimensional solid objects could be. Why is the Universe apparently so simple?

In this section: the large-scale shape of the Universe could be very complex. Even if the Universe is homogeneous, it could be anisotropic: different in different directions.

► Anisotropic expansion is different from what an observer would see if the extra speed of “expansion” were really caused by the observer’s own motion, for then in one direction galaxies would be receding more rapidly than in perpendicular directions, while in the opposite direction galaxies would be receding more slowly.

► Anisotropic expansion also challenges Newtonian cosmology: if the universe is infinite, should the gravitational field be calculated by dividing space into spheres rather than, say, into ellipsoids with the same shape as the expansion velocity? Newtonian gravity offers no unique resolution, while general relativity gives unique answers.

In this section: observations give no evidence that the Universe has any other large-scale geometry than the simplest: a homogeneous and isotropic space.

In this section: we learn how to describe and measure the curvature of a homogeneous and isotropic space.

in a particular direction as in any perpendicular direction. Notice that, because this is an expansion, it looks the same if one looks in one direction or in exactly the opposite direction (which means turning 180° around). So the rapid expansion we are imagining occurs in both directions along a particular line.

Now, just having a higher expansion rate in one direction does not destroy the *homogeneity* of the universe: no matter where the observer is, the expansion in that particular direction would be twice as fast as in perpendicular directions. Just imagine a sheet of rubber as a two-dimensional universe model, and let the sheet be stretched in only one direction. It can remain homogeneous – the same everywhere – even as it expands. Because not all directions are the same, the expansion is not isotropic (recall the discussion of isotropy in Chapter 7). We say that this kind of universe model is homogeneous but *anisotropic*.

This sort of expansion could occur in a Newtonian universe too: if we start the universe off with such an expansion, then Newtonian gravity will keep it expanding in this anisotropic way. But in an Einstein universe model, the anisotropic expansion changes the gravitational field itself, and such models can differ dramatically from Newtonian ones.

So in principle, cosmology *could* be much more complicated than the Newtonian universe models we have studied so far.

... but in fact it is simple (fortunately!)

But the observational evidence all supports the simple cosmologies:

When we look for evidence of this kind of anisotropy in the real Universe, we find none.

In particular, the cosmic microwave background radiation does not show any big systematic effects of this kind: once we have removed the Doppler effect of our own motion, its temperature is the same in all directions to a high accuracy. If the Universe were expanding anisotropically, we would expect to see one temperature along the direction of the more rapid expansion (in both directions along this line), and a different one in perpendicular directions. The deviations that we do measure seem to be random: there is no large effect along one line.

This is the **homogeneity/isotropy problem**: out of all the possible kinds of universes we might have found ourselves in, it seems puzzling that ours is so nearly homogeneous and isotropic: why has Nature provided us with such a simple arena to play in?

The idea of inflation, which we shall study in Chapter 27, is an attempt to provide an answer to this question, among others.

Gravity is geometry: what is the geometry of the Universe?

Einstein described gravity in terms of geometry. So when we look for model Einstein universes similar to the Newtonian ones we met in Chapter 24, we need to look for *geometries* for three-dimensional space that embody the remarkable homogeneity and isotropy that we see around us. When we do, we find that things are much simpler than we might have expected.

There are in fact only *three* possible kinds of homogeneous and isotropic models. They are commonly called the closed, open, and flat universe models. These three cosmological models were first discovered by the Russian physicist and meteorologist Aleksandr Friedmann (1888–1925). In Investigation 26.2 on page 389 we look at their geometry in some detail. But it is important here to understand why there are only three, and what they look like.

We are all familiar with at least one three-dimensional space that is isotropic and homogeneous: the standard Euclidean space that we grew up thinking we lived in! This flat model universe is one of Friedmann's geometries. What about curved spaces: are there any that are still homogeneous and isotropic?

Now, there are many ways to distinguish a flat space from one that is curved, but if the space is isotropic, so that nothing changes from one direction to another, then things are fairly simple. Notice first that a space that is fully isotropic has to be homogeneous. The reason is that, if it were *not* homogeneous, then there would be at least one place where things were different from other places (say, a "bump" somewhere). Now, if we were to stand anywhere else in the space and look around ourselves, we would see that space was different in one direction (looking towards the "bump") than in other directions: the space could not be isotropic about our location. For the space to be isotropic about all its points it must also be homogeneous.

Now, if a three-dimensional space is isotropic, then we can be sure that we can draw a perfect sphere about any point in it: since all directions are the same, then we just form a surface from all the points whose distance from the central point is constant. We call this distance the *radius* of the sphere. The sphere will be identical to a sphere in Euclidean space, and in particular we can draw a great circle on it, say its equator. The length of this circle is called the *circumference* of the sphere.

All this may seem trivial, just elementary geometry, but it turns out that such spheres give us a very simple measure of the curvature of the space we are in: just draw a sphere around any point and take the ratio of its circumference to its radius. In a Euclidean space, this ratio is of course just 2π . The converse is also true: if we are in an isotropic space and we find that this ratio is exactly 2π for one sphere drawn about one point, then it will be exactly 2π for a sphere of the same radius drawn about any other point, and by constructing all of these spheres we can build up the space and find that it must be Euclidean. How to do this is shown in Investigation 26.1 on the next page.

More interestingly, suppose we have an isotropic space and we measure the ratio of circumference to radius of one particular sphere and find that it is *smaller* than 2π . Then that space must be curved. Since the space is homogeneous, spheres of that size drawn around other points will have the same ratio, and (again as we show in Investigation 26.1) we can construct the space just from these spheres. In particular, the circumferences of other spheres of different sizes will all be determined by the properties of the first sphere we chose.

The sphere we chose first is not the only one that would generate this space: any other sphere in this space would have worked just as well. Spheres of different radii may have different ratios of circumference to radius, so this ratio for an arbitrary sphere does not characterize the curvature of the space. One way of defining the space and its curvature is to look at a *particular* sphere, say the one that has a radius of 1 m. Its circumference is a single number that completely defines the space: there is one and only one space for which a 1 m-radius sphere has that circumference.

We have learned that a homogeneous and isotropic space is determined by giving just *one* number.

If that circumference is 2π m, then the space will be a three-dimensional Euclidean space.

But what if the circumference of the 1 m sphere is *less* than 2π m – is that really possible? How can one imagine a space that has circles whose radii are too large for their circumferences? The answer is that it is actually very easy to visualize such a space if we go down to two dimensions and try to find a two-dimensional surface

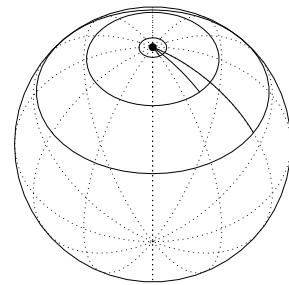


Figure 26.1. When we do geometry on the sphere, we find that circles have smaller circumferences than they do in the flat Euclidean plane. Suppose that the sphere shown here has a radius of 1 m. Three circles centered on the pole are displayed. The smallest circle has a radius, measured along an arc drawn on the sphere down from the pole, of 0.1 m. A measurement of the length of its circumference would in fact give 0.62727 m, so their ratio is 6.2727, which is only 0.9983 times 2π . For the middle circle, whose radius is 0.5 m, the ratio is 0.9589. And for the largest circle shown, of radius 1 m, the ratio is 0.8415. The key point is that we are talking about the geometry of the sphere, so the radius of a circle must be measured along the sphere, not along a line that extends out into the surrounding three-dimensional Euclidean space.

Investigation 26.1. The sphere, and only the sphere: a two-dimensional universe

Our goal here is to see why there is only one kind of two-dimensional homogeneous and isotropic space in which circles have smaller circumferences than 2π times their radii. We shall do this by construction, by making a sphere out of the circles.

We start from the premise that our space contains at least one circle whose circumference is a smaller multiple of its radius than it would be in flat space. We begin by choosing a size for this circle and drawing it in three-dimensional Euclidean space.

Our next step is to choose the location of its center, by which we mean the point on the two-dimensional sphere that will be its center. If we place this center in the plane defined by the circle, then the radius will be the circumference divided by 2π , which is exactly what we have in flat space because the plane containing the circle is flat. To make a sphere we need to give the circle a larger radius for its circumference. To do this we simply move the point that represents its center out of the plane of the circle, along a line perpendicular to that plane. Then our circle-plus-center looks like the largest circle in Figure 26.1 on the preceding page. Our circle is shown in the upper left in Figure 26.2. Its center is not shown in this diagram. We will see that choosing the first circle and the location of its center fully defines the 2-sphere.

What we cannot yet do is to draw the circle's radius in the 2-sphere, because that would require us to know the shape of the surface between the center and the circle. We have to "build" this space, so all we can do is start with a center displaced from the expected position. We can say something about the radius, however. The two directions along the circle must be equivalent, or the space would not be isotropic. Therefore, the radius must intersect the circle at a right angle. Any other angle would allow one to distinguish one direction from the other around the circle. Since this applies to smaller circles as well, the radial curve must lie in a plane perpendicular to the plane of the circle itself, and containing the displaced center.

The key to the construction of the sphere from this first step is a simple sequence of operations. Consider any point on the circle itself. This is a point of the space we are constructing, and that space is homogeneous. Therefore, if we move to this new point and look at the space, it must look the same. In particular, there must be another circle-plus-center, identical to the one we have just defined, in the space, with its *center* at this new point. The key to the construction is to determine how this second circle-plus-center fits with the first one. Its center is the new point, but how is it oriented?

The first part of the answer is that the second circle has to pass through the center of the first circle. This is because the points on the circle all have the same distance from the central point, and are indeed defined by this distance. Since the new point has this distance from the first center, then the first center must be on the equivalent circle drawn around the new point.

The second part of the answer follows from this: the new and old circles must *share* the same radial curve. The reason is that the radius must be the shortest distance within the space between the center and the circle, and so it must be the same curve whether we regard one point or the other as being at the center or on the circle. Now, we have already seen that this radial line must be perpendicular to the old circle. It must therefore also be perpendicular to the new one, and so both the old and new circles lie in planes perpendicular to the radius. This fixes the orientation of the new circle: given its center on the first circle, it must pass through the first center

and lie in a plane perpendicular to the radial curve. We construct the new circle by rigidly moving the first one until its center is on the new central point, and we tilt it until the new circle passes through the old center, keeping it perpendicular to the plane that we just identified.

The old and new circles are shown in the upper right in Figure 26.2. Notice that the two circles intersect at two points. This is good: we want them to form parts of the same surface, so they should certainly not pass over or under one another.

Now, the space we are constructing is also isotropic, so the same thing must happen at any other point on the first circle. This leads to a set of new circle-plus-center constructions distributed around the original circle, all passing through the first central point. We show in the bottom left in Figure 26.2 the members of this family that are separated by 1 radian around the first circle. By construction, these all intersect at the point that we chose as the center of our first circle, so the location of this point in our space is now clear from this diagram.

Then one can further build up the space by doing the same on each of these circles, allowing them to spawn more circles by taking their points to define new centers. The result of allowing one of the secondary circles to spawn more is shown in the lower right of Figure 26.2. It should be clear by now that we are filling in the ordinary 2-sphere. The radius of the sphere is determined by the displacement of the center above the plane of the circle that we adopted for the first circle: the smaller the displacement, the larger the sphere, and the closer to flat space the space is. But for any non-zero displacement, the construction will give a 2-sphere.

If by chance we had chosen a circle-plus-center object that turned out to span exactly 90° on the 2-sphere, then the circles would keep repeating and would not fill in the whole sphere, but with the exception of a set of such special objects, the repeating circle-plus-center objects will eventually pass through every point of the sphere.

This construction does not entirely rigorously show that the 2-sphere is the only two-dimensional homogeneous and isotropic space containing circles whose circumferences are smaller than 2π times their radii. Strictly speaking, it is only part of a rigorous proof. The rest of the proof must address the question of whether we were justified in assuming in the first place that our circle-plus-center should be drawn in three-dimensional Euclidean space. Maybe we would get another kind of surface, not a 2-sphere, if we started in a different space, say five-dimensional Euclidean space.

It is not hard to show that there is nothing new in five-dimensional Euclidean space. But there is something new if we start in the Minkowski spacetime of special relativity, as we describe in Investigation 26.2 on page 389. In such a space it is possible to draw circles whose radii are *shorter* than the circumference divided by 2π . By combining those circles in the way we have done here, one constructs, not a sphere, but a hyperboloid.

There are thus just three possibilities for constructing a homogeneous, isotropic, two-dimensional space: construct a sphere in Euclidean space out of circles with radii that are too large, construct a hyperboloid in Minkowski spacetime out of circles with radii that are too small, or construct a plane out of circles with radii that are just right.

Goldilocks would be pleased!

where the radii of circles are extra large compared to the circumferences. We have only to look at an ordinary sphere, as in Figure 26.1 on the preceding page. There we see three circles centered on the same point. Their radii are arcs drawn on the sphere, and from this it is easy to see that as the arcs get longer the circles do not grow in circumference as rapidly as they do in flat space.

Now, the sphere is a two-dimensional, homogeneous and isotropic space. Every such space in which circumferences are "too small" is a sphere of some size. The reason is that one tiny patch of the surface determines the whole surface, by homogeneity and isotropy. So if, as one increases the size of a circle from nothing

to a tiny amount, the circumference begins to lag more and more behind 2π times the radius, and one only needs to know how serious the lag is: in one centimeter, does the circumference fall below $2\pi \times 1\text{ cm}$ by 1%, by 0.1%, by 0.01%, or ...? This fractional lag determines the radius of the sphere: the smaller the lag, the more nearly flat the space is, so the larger the radius of the two-dimensional sphere. The uniqueness of the sphere can be shown by explicitly constructing the sphere from the circles. We show how to do this in Investigation 26.1 and Figure 26.2.

Friedmann's model universes

We have already seen that one of the model universes in general relativity will be three-dimensional Euclidean space, the obvious homogeneous and isotropic three-space. We have also seen that in two dimensions the sphere is likewise homogeneous and isotropic. The generalization of the two-dimensional spherical surface to 3 dimensions is an important geometry that we call the three-sphere.

The 3-sphere is defined as the set of all points that have the same distance from a central point in four-dimensional Euclidean space. Now, it is not easy to visualize four-dimensional Euclidean space, so I don't recommend trying. The properties of the three-sphere are very like those of the ordinary sphere (the 2-sphere). In particular, it is the only homogeneous and isotropic three-space in which spheres have circumferences that are "too small". Therefore, the three-sphere is also one of Friedmann's model universe geometries. It is usually called the because it has finite size.

Before asking about the third kind of model universe, I want to warn the reader about one pitfall in studying Figure 26.1 on page 385. It may seem that the excess radius is a cheat, that the circles would look like normal circles if we looked instead at their radii, not running along the sphere back to the pole, but in a plane slicing through the sphere and containing the circle. Since that plane is flat, the ratio of circumference to this kind of radius is still 2π . This is of course true, but irrelevant. Figure 26.1 is meant to illustrate a property of the *intrinsic* geometry of the sphere, i.e. what we would measure about circles if we were little ants confined to the surface of the sphere, only able to lay out lines and take measurements on the surface. We must confine ourselves to the surface in order to make a good analogy with the three-dimensional case that we are interested in: when we generalize to a three-sphere, then there is no physical way to slice through the circle to get it to be flat: one would have to slice into a fourth, unphysical, dimension to do this. In all three dimensions of a three-sphere, circumferences grow more slowly with radius than in Euclidean space.

And what about the opposite case, where circumferences grow more rapidly with radius than in Euclidean space? By the same reasoning, there is only one kind of three-dimensional geometry that has this property. It is harder to visualize this geometry, however, because an analogous two-dimensional surface cannot be drawn in

In this section: there are three types of homogeneous and isotropic Universe model: spherical, flat, and hyperboloidal. We learn how to construct them explicitly.

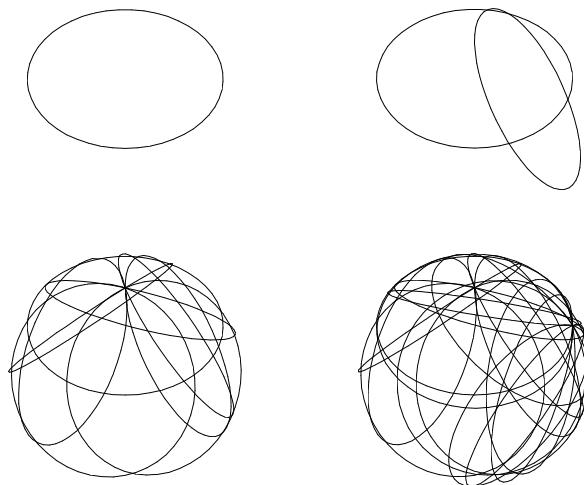


Figure 26.2. A construction to show that the sphere is the only two-dimensional surface that is homogeneous, isotropic, and contains circles whose radii are larger than their circumferences divided by 2π . The four stages of construction are described in Investigation 26.1.

>This is, indeed, the situation we face if we do civil engineering on a large scale on the Earth: the way large road networks mesh is significantly non-Euclidean, and one can't make a faithful roadmap on a flat piece of paper if the area described is more than a few hundred kilometers in size.

► If the Universe we live in is really described by a three-sphere or 3-hyperboloid, is there any physical reality to the fourth dimension of a Euclidean or Minkowskian space in which the sphere is drawn? Is that dimension "really" there? There is no logical necessity for it to exist, since all our experiments are performed in three dimensions, but it might be philosophically pleasing if it did. Maybe this dimension might actually be one of the dimensions in the higher-dimensional spaces of string theory (see Chapter 27).

three-dimensional Euclidean space like the 2-sphere can be. We describe in Investigation 26.2 how the geometry can be drawn as a surface in the Minkowski spacetime of special relativity. In this space, it looks like a **hyperboloid** rather than a sphere, and so we call it the hyperbolic model universe, or the *open model universe*. It is infinite in extent.

Friedmann did more than just characterize the geometries. He showed how each geometry is linked to the properties of the matter that create it. Recall our discussion of spatial curvature, Equation 19.2 on page 244, which asserts that in Einstein's equations, spatial curvature is produced by the combination $\rho - p/c^2$, at least for isotropic pressure. We wrote this down in the weak-field case, assuming the geometry was close to flat spacetime. In the case of cosmology the geometry has an added feature, namely the expansion. Friedman's geometries are spheres or hyperboloids at any one particular time, but they expand. The result of this is to modify the expression and replace the pressure by a term involving the Hubble expansion parameter:

$$\text{curvature of space} \propto \rho - \frac{3}{8\pi G}H^2. \quad (26.1)$$

Effectively, the pressure is replaced by an effective pressure equal to $3H^2/8\pi G$ when determining the spatial curvature. This applies even when there is real pressure in the cosmology: the sum of the pressure and any kinetic stresses (ram pressure as in Chapter 19) is replaced by the expansion term in H^2 .

Now, if we use Equation 24.10 on page 361 to replace H^2 by the critical density ρ_c we get the simpler equation

$$\text{curvature of space} \propto \rho - \rho_c. \quad (26.2)$$

A universe that has just the critical density will have zero curvature, so it will be spatially flat; one that has more than the critical density will have positive curvature, which is to say that it will be the closed three-sphere; and one that has less than the critical density will be the open, hyperboloidal model.

The large-scale geometry of the three models is very different: the sphere is a finite space, the hyperboloid is infinite, and the flat model is, well, flat. It should not be a surprise that it is not possible for one model to transform itself into another at some time: the geometrical class is preserved during the evolution of the cosmology.

If the density is critical at one moment (so that the Universe is flat) then it remains critical for all time.

This is not achieved by some mysterious mechanism; all that happens is that the acceleration equation (Equation 24.12 on page 362) insures that the Hubble constant changes with time at just the right rate to keep the critical density equal to the current mass density.

For matter-dominated cosmologies, the critical density is also related to the escape speed.

Closed matter-dominated models will re-collapse, while open models expand forever.

If pressure is significant, however, this simple link between curvature and the future of the Universe is broken. A model could have more than the critical mass density, and so be closed; and yet in principle it could have enough pressure to make its active gravitational mass smaller, or even negative, so that it expands forever. We saw in Chapter 25 that a universe dominated by a cosmological constant will expand exponentially fast, so it never reaches a constant speed.

Investigation 26.2. The three Universes of Friedmann

In this investigation, we explore the nature of the three Friedmann geometries for the Universe. Since the shape of the Universe outside our past light-cone does not affect us, we can make any assumption we want about it. It simplifies the discussion, and fits the Copernican principle, if we assume that it is exactly like the part of the Universe we do observe. But we must bear in mind that this is only a matter of convenience.

We therefore ask what kinds of three-dimensional spaces are homogeneous and isotropic everywhere. Clearly, flat Euclidean space fits our requirement: no matter where we are, the geometry is the same as anywhere else, and in all directions. This is the simplest of the universe models.

The second kind of homogeneous and isotropic three-dimensional space is the three-sphere. It is a generalization of the usual sphere, which is called the 2-sphere because its surface is a two-dimensional (curved) space. (The 1-sphere is simply a circle, of course!) A perfect sphere is the same everywhere: the geometry has no bumps or defects to tell one where one is, and turning the sphere around does not change anything either. The 2-sphere is a two-dimensional isotropic homogeneous space.

The three-sphere is the same, extended to one more spatial dimension. Its formal definition is the set of all points in four-dimensional *Euclidean* space that are the same distance from a given point. Most people find it hard to visualize such a thing. I recommend not trying to, but instead trying to understand its properties by generalizing from 1- and 2-spheres.

Like a 2-sphere, which has a finite area, the three-sphere has a finite *volume*, but nevertheless a curve drawn in it never encounters an edge: curves just keep circling around and around the space. A three-sphere it is what mathematicians call a finite and *closed* space, which means it is finite but has no boundary.

Now, one can measure the radius of a sphere by measuring the way the circumferences of circles increase with radius, as in Figure 26.1 on page 385. We constructed a 2-sphere from this information in Investigation 26.1 on page 386. This is an important point: if we walk outwards in our own Universe and find that the circumference is not increasing as rapidly as we would expect in flat space, then it must be a spherical Friedmann universe, and we can even measure its radius! Once we have measured the size of the three-sphere by examining circles in some small region, the rest of the geometry has to wrap itself up into a three-sphere of this size, or else it must be inhomogeneous.

The other possible universe model is the one where the circumference of a circle increases *more* rapidly with radius than in flat space. This space can be described as a subset of points four-dimensional space *Minkowski spacetime*, which we met in Chapter 17. The third three-space has a definition that is closely analogous to that of the three-sphere: it has the same geometry as the set of all points that are at a constant *timelike spacetime-interval* from a given point in Minkowski spacetime.

What this means is the following. Choose any event in Minkowski spacetime to be the origin, $t = x = y = z = 0$. The timelike spacetime-interval between the origin and another event is just the

time ticked on a clock that travels at a constant speed from the event at the origin to the other event. Another clock, traveling at a different speed and in a different direction, will reach a different event after the same time has elapsed on it. The set of all such events, reached by all possible clocks traveling at less than the speed of light, is the third space that is a possible Universe model. Because the spacetime-interval is given by Equation 18.6 on page 230, the equation for this space is

$$c^2 t^2 - x^2 - y^2 - z^2 = \text{const},$$

which is the equation of a hyperboloid in Minkowski spacetime.

This space is homogeneous because all observers agree that all possible clocks started at a particular event and arrived at the surface with the same *proper* time. This is observer-independent. Different observers regard different clocks as being at rest, so they would place the origin of coordinates in different places, but the space would look the same to them.

How do circles behave in this space? In this case, as a circle is enlarged, its radius increases along the hyperbola. Now, the hyperbola is getting closer and closer to a lightlike direction as we move outwards from the center of the coordinates. This means that the proper length of the radial curve does not increase very much as the circle is enlarged: lightlike lines have zero proper length. The result is just the opposite of the case for the sphere. As the circle increases in circumference, the radius fails to increase much, and the ratio of circumference to radius is *larger* than in flat space.

Just as for the sphere, even a small circle is enough to tell us the amount of curvature. Therefore, once we have determined how the curvature affects one circle, we have determined the geometry of the space. Mathematicians call this sort of curvature *negative*.

The hyperbolic three-space is still a true space, despite the fact that we have constructed it in a spacetime. Any two points within it are separated from one another by a spacelike spacetime-interval because they happen simultaneously to an observer whose velocity is the average of the velocities of the clocks that reach those points.

This space is not closed. With respect to a given experimenter, clocks that travel very close to the speed of light take longer and longer to tick the given time-interval. This is just the time dilation of special relativity, and it means that these clocks define points on the space that get further and further away from the origin. The hyperboloid just keeps going in Minkowski spacetime. It is an *open* space.

We have constructed three model universe spaces. And this is all there is: our construction leaves no room for any other homogeneous and isotropic spaces.

Cosmologists therefore speak of three possible Universe models: the flat, closed (spherical), and open (hyperboloidal) models. These names refer to their overall structure, which depends on the regions outside our past light-cone. However, their local geometry also reflects their shape, and this is measurable. We must always bear in mind that the only observable features of these models are their sections inside our past light-cone.

Notice that we can determine the geometry of the Universe by measuring the total mass density of the Universe in our neighborhood and the Hubble parameter. We will see in the next chapter that, when we do this for the observed Universe, we find that we appear to live in a nearly flat cosmology. The density of matter plus that of the dark energy is enough to reach the critical density, within observational uncertainties. Flatness is another prediction made by inflationary models of the early universe, as we will see in the next chapter.

However, it is important always to bear in mind that we can ever only know about a finite portion of the Universe, and our conclusions only apply to what we can see within our past light-cone. Beyond our particle horizon, the geometry might be very different. And if it is – so that the Universe is not homogeneous – then its future evolution will be different from what we would predict as well.

What the Universe looks like

In this section: the curvature of space produces larger angular diameters in distant objects than in flat space. This could be observed by astronomers.

►The angular diameter is the angle on the sky that the object appears to occupy in our observation. This concept was first introduced in Chapter 5.

Figure 26.3. The curvature of the Universe can dramatically affect the apparent angular size of an object. Two galaxies of identical physical size (shaded ovals) are at different distances from the observer in a closed spherical universe. Since we only want to see the effect of spatial curvature, we suppose the sphere is not expanding. Light travels along great circles, so the light rays from the more distant object to the observer (dashed lines) can arrive with a wider angle than those from the nearer. This would make the more distant object look larger than the nearer.

How do we measure the curvature of the Universe? The simplest way is to try to measure the total mass density and the Hubble constant, and take the difference between the observed density and the inferred critical density, as in Equation 26.2 on page 388. This is an indirect determination of the curvature, since we are measuring the quantities that produce the curvature, rather than the curvature directly. But it has the great advantage that astronomers are already measuring the Hubble constant and the density for many reasons. The best estimates today suggest that the curvature is close to zero, that we live in a nearly flat cosmology. We will look more at these measurements in the next chapter.

Another approach might be to look for direct evidence of curvature. Distant galaxies and quasars are separated from us by a curved spacetime. There must be evidence of this curvature in observations of them.

The kind of evidence that one might look for is not hard to imagine. It follows directly from our way of defining the curvature of the Friedmann universe models: a geometry is curved if the circumference of a circle divided by its radius does not equal 2π . Now, if we observe a distant object whose size is known, then if we measure its *angular diameter*, then the circumference of a circle at the distance of an object is just its size divided by its angular diameter. We can measure the radius of the circle if we know how far away the object is, which is given by its redshift.

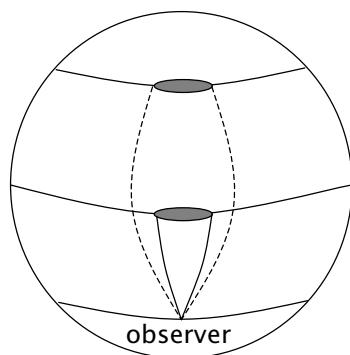


Figure 26.3 illustrates an extreme example of what can happen in a curved universe model. In the figure, the Universe is the closed three-sphere, and two objects (galaxies) of the same size are observed at different distances. Since the size of circles around the sphere begins to decrease as you go further and further away, the fraction of a full circle occupied by the more distant galaxy is much larger than that of the nearer, so its angular size is actually *larger* than that of the nearby galaxy, despite being further away. The general rule is that if the universe model is closed then angular diameters decrease less rapidly with distance than in a flat universe, and if the model is open they decrease more rapidly.

Astronomical observations that would reveal such an effect require a **standard meter stick**, an object whose physical size is the same in the early Universe as it is today. Like standard candles, a class of objects must be found whose size is not affected by evolution, or at least whose evolution in time is understood. A number of possible objects are currently the subject of investigation, including the angular sizes of the fluctuations in the microwave background radiation. It is possible that in the near future astronomers will have an independent check on the curvature of the Universe by measuring angular diameter evolution.

One theory actually predicts the curvature of the Universe: if the very early Universe underwent a period of inflation, as many astrophysicists now believe, then it should have stretched space nearly flat. Inflation is one of the subjects at the frontiers of physics and cosmology, to which we now turn.

Ask the Universe: cosmic questions at the frontiers of gravity

The study of cosmology presents today's physicists with the biggest challenges to their understanding of gravity and of fundamental physics in general. Both on theoretical and on observational grounds, it seems that we will not be able to understand cosmology well until we understand physics better than we do today. But it also seems that cosmology could provide us with the keys to that deeper understanding of physics.

The biggest gap in physics is quantum gravity: we do not yet possess a consistent way of representing gravity as a quantum theory. There is no uncertainty principle in general relativity, no quantization of gravitational effects, no need to use probabilities in making predictions about the outcome of gravitational experiments. This seems inconsistent with the fact that all material systems that create gravity are quantum systems: if we can't say exactly where an electron is, how can we say exactly where its gravitational field is?

Many physicists believe that the way to quantize gravity is to unify it with the other forces of nature in a single theory in which electromagnetism and gravity are just different members of a single family of forces, and in which the unity among these forces only becomes apparent at very high energies, near the Planck energy. One would expect such a theory to predict new phenomena at these high energies. The only places where we know such energies have been met in the history of the Universe are (1) inside black holes, and (2) at the Big Bang. Phenomena inside black holes are hidden, but the Big Bang is very visible.

By a combination of theory, experiment, and observation, physicists and astronomers hope to use the Universe as a laboratory to make big advances in physics. Cosmology is a hunting ground for clues to the ultimate unification of the physical forces.

Fortunately, there are many clues. We have met a number of them in passing during earlier discussions. Here is my personal list of cosmic puzzles whose solutions have the potential to revolutionize our understanding of physics.

- *Clue.* The Universe on the large scale is *homogeneous* and *isotropic*. Regions that, in the conventional Big Bang model, have not yet had enough time to have communicated with each other (see Figure 27.1 on page 393) are nevertheless very similar. They have the same density of matter, the same numbers and types of galaxies, the same degree of clustering; the microwave background radiation has the same temperature in all directions around us to a few parts in 10^6 ! How was this arranged?
- *Clue.* Galaxies could not have formed without *dark matter* seeds. Experimental searches for the dark matter may soon show what the cold component is. If it turns out to be a new elementary particle, then its identity will be a clue

In this chapter: we confront the limits of modern physics with puzzles and clues from cosmology. They have to do with the large-scale properties of the Universe, the formation of galaxies, and even the formation of life. The next big step in theoretical physics will be the unification of gravity with the other forces. The resulting theory should be able to address the questions we ask here, and go beyond them. It should clarify quantum theory, and even tell us something new about time itself.

>The unified theory could also predict new phenomena at lower energies, but none have been noticed in experiments. Some physicists have recently suggested that they might modify Newtonian gravity, making it stronger on distance-scales shorter than some characteristic length, which could be as large as 1 mm!

>The image under the text on this page is a *simulation* of the kind of data expected from NASA's MAP satellite, which began observation of the cosmic microwave background shortly before the completion of this chapter (2002). It will provide the most detailed map to date of the microwave background's irregularities. (Compare with Figure 25.4 on page 381.) These in turn will give physicists their best measures so far of the conditions in the very early Universe. Even higher-resolution data should come from the Planck mission, planned for launch by ESA by 2007. Courtesy MAP Science Team and NASA/GSFC.

to new kinds of physics. There are plenty of candidates for this new physics already, but scientists need experimental or observational data to tell them which ideas are right.

- *Clue.* When astronomers make their best estimate of the total mass density of the Universe, adding in the dark matter and dark energy densities, they find it *equals the critical density* as defined in Equation 24.10 on page 361. This is a very special value, because if the Universe is critical at one time, it remains critical for all time. Many physicists feel that this should have an explanation.
- *Clue.* Observations of the cosmic microwave background strongly support the idea of *inflation*, that the Universe underwent a very early phase of enormously rapid expansion, which was driven by dark energy with a negative pressure, like a temporarily large cosmological constant. The cause of inflation is shrouded in our ignorance about physics at the highest energies, but it is already clear that many fundamental processes can mimic a cosmological constant.
- *Clue.* Observations of the expansion of the Universe seem to show that the Universe has again entered a phase of *accelerated expansion* with a much smaller dark energy. This could be a remnant of the earlier inflationary phase, or a new physical field, or a permanent cosmological constant (or all three!).
- *Clue.* Theories of high-energy physics suggest that the Universe may contain *cosmic strings*, long concentrations of dark energy, thinner than any elementary particle. Cosmic strings do not curve time but they do curve space, and they could be detected by gravitational lensing.
- *Clue.* Observations of the *highest-energy cosmic rays* have shown that the Earth is struck by about one cosmic ray with an energy larger than 10^{20} eV each second. This is a tiny flux of particles that have incredibly energy, some 10^8 times greater than physicists can produce in particle accelerators. The origin, and even the nature, of these cosmic-ray particles is a complete mystery. Maybe their sources are dark and represent new physics, or maybe the particles themselves are new.
- *Clue.* The Universe would not have produced human beings if the laws of physics did not have some very special properties, including some apparent *coincidences among the fundamental constants of nature*. Some scientists see in these accidents a role for God, as the creator of the improbable machinery that led to life. But many others look for explanations within physics. We will discuss several mystifying coincidences below.

▷For comparison with the flux of these cosmic rays, recall that in Chapter 11 we saw that in each second ten billion neutrinos of much lower energy pass through your body alone!

▷Could the coincidences be explained by selection from a large set? For example, does the Universe have many Big Bangs in different places, each beginning with a different set of randomly chosen constants, so that some are guaranteed to allow people to evolve and ask these questions? (The British astrophysicist Martin Rees (b. 1942) has called such a universe a “multiverse”.) Many physicists treat such speculations seriously, and they hope that quantum gravity will provide serious answers.

We will go through this list of puzzles in this chapter, weaving these challenges into a larger discussion of quantum gravity and the prospects for a unified theory of all the forces of Nature.

Unlike in previous chapters, here we are at the frontiers. Physicists’ perspectives on what is puzzling, important, or fundamental change rapidly here. Even by the time you read this, some of the puzzles on our list may have been resolved; others may have been rendered irrelevant to fundamental physics; new ones might join the list. Progress, stimulated by new observations and new theoretical speculations, is sure to be rapid but unpredictable. But don’t underestimate the difficulty of arriving at a full understanding of the physics of the Universe. It is work for a generation of physicists. Or more.

The puzzle of the slightly lumpy Universe

The Big Bang model of cosmology provides a framework for thinking about the history and future of the Universe, and we have seen that this framework provides simple cosmological models that seem to fit the observed facts very successfully. But the simplicity of these models raises two big questions which the Big Bang model does not answer.

Why is the Universe so smooth on large scales, so homogeneous?

Given that it is smooth, why is it not even smoother? Why did it have enough initial irregularity on small distance-scales to form the stars and galaxies we see?

We have seen in earlier chapters that the homogeneous Friedmann model is a good model of the Universe. At the time of nucleosynthesis, when helium was being produced, the Universe was remarkably smooth on the very smallest scales: the ratios of the different isotopes show no evidence of the slight changes that would have been caused by significant inhomogeneity.

Dramatically, there are many measures that show us that the Universe was homogeneous at very early times, so early that the regions we compare would not have had time to communicate with each other in the standard Big Bang model. The helium abundance is the same in different directions; the numbers and types of galaxies at very high redshift are the same in opposite directions; and the microwave background temperature is the same to a few parts in 10^6 all over the sky!

Provided we believe we are not in a special place in the Universe from which it just happens to look homogeneous (this is the Copernican principle, introduced in Chapter 24), then this homogeneity poses a problem, a problem of *how*. The two regions producing helium could not have influenced one another physically in order to insure that they made the same amount of helium (by arranging to have the same density and temperature, for example), so *how* have they arranged to be so alike?

If the standard Big Bang is right, then the large-scale homogeneity we observe seems to be accidental. It requires that the initial conditions for different parts of the Universe were the same at the Big Bang. A messy, random initial start to the Universe could not have produced the Universe we see.

This conclusion is disturbing from the point of view of the Copernican principle, since it means that the Universe itself is very special: of all the kinds of universes that one could imagine, the one we have is exceptionally homogeneous.

Inflation modifies the standard Big Bang picture to offer an explanation of the homogeneity. Inflation proposes that there was a very early period dominated by dark energy that acted like a temporary cosmological “constant”, driving a very rapid (exponential) expansion. The result is that large regions today have actually expanded from tiny regions just before the onset of the inflationary expansion, regions that were small enough to have become smooth in the very short time between the Big Bang and the beginning of inflation.

If inflation lasted long enough, with a strong enough acceleration, then everything we see today was once inside a tiny region. The smoothness we see on short and long scales could have been achieved by ordinary physical processes before inflation, even if the Universe initially had been very messy and random.

In this section: the Universe is smooth on the large scale but lumpy enough to make galaxies. This combination is very special. Inflation provides explanations.

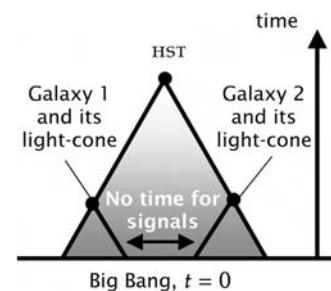


Figure 27.1. Two galaxies, observed by astronomers using the Hubble Space Telescope (HST), lie in opposite directions to one another, at such a great distance that they are seen when the Universe was much younger. They are so far apart that there has not been enough time for them to communicate, to exchange signals: their past light-cones reach the Big Bang before they intersect. This is the conventional picture of the Big Bang without inflation, and shows why the homogeneity of the Universe – the fact that the galaxies and their neighborhoods look so similar – is a difficulty for the standard Big Bang model. It requires that the Big Bang should have started in a very similar way in disconnected places.

► The absence of an explanation for homogeneity or clumpiness is, of course, fundamentally a philosophical difficulty rather than a strictly scientific one, since there is only one Universe, and a scientist can only measure what it is like, not how it might have been if things had been different at the Big Bang. Cosmology is not an experimental science: one cannot create new universes under controlled conditions and explore what happens to them!

In fact, for some physicists homogeneity has a religious implication, proving that there was a design in the Big Bang, that the initial explosion of the Universe was carefully wrought by a Creator.

But other physicists feel that it is now possible to look seriously for purely scientific explanations of the structure of the Universe. For them, the philosophical difficulty has been fruitful scientifically, because it has led them to the idea of inflation.

In this section: Einstein regretted introducing the cosmological constant, but today many scientists believe it exists in some form.

But wait: this explains the homogeneity of the Universe, but what about its lumpiness? The uniformity certainly breaks down on small length-scales, because galaxies and stars have formed. Doesn't inflation smooth things out too much?

Remarkably, inflation offers an explanation of the lumpiness to. The expansion due to inflation had a side-effect: it created density irregularities through the amplification of quantum fluctuations that existed before inflation, the same kind of fluctuations that we described in our discussion of black-body radiation from black holes in Chapter 21. We will see how such tiny effects can give rise to the formation of massive galaxies in our discussion of inflation below.

Inflation's explanation of galaxy formation is the critical argument that makes scientists take the idea seriously. It produces the right amount of inhomogeneity, on the right length-scales. Inflation promises to solve both the smoothness problem *and* the not-so-smoothness problem.

Inflation has given scientists the confidence that they can solve the mysteries of cosmology with scientific methods. They remain strong in their belief that the Universe is understandable.

The combination of Einstein's law for gravity and modern thinking about the fundamental laws of physics may have enough power to penetrate through barriers that physicists formerly regarded as absolute, and to answer questions that physicists once would not have dared to ask. The answers, if they come, will come from the Universe itself.

It is time now to begin our study of inflation. The first step is to look at the cosmological constant in its cosmological setting.

Einstein's "big blunder"

When Einstein invented general relativity, Hubble had not yet discovered the expansion of the Universe, and the general opinion of astronomers of the day was that the Universe was static, that it had always existed in its present state. This was, of course, fundamentally because they were only looking at stars in our Galaxy. But it was also conditioned by the Judeo-Christian philosophical and religious beliefs shared by the scientists who developed physics in the nineteenth and early twentieth centuries.

Their society widely believed that the Universe was created at a finite time in the past, and in a "perfect" state, in just the same condition as they observed it. The idea that the Universe was dynamical – changing with time, evolving – would have been uncomfortable, in the same way that the idea that life was evolving was uncomfortable to many people at that time, scientists included. So, in the absence of evidence to the contrary, nineteenth century astronomers took the conservative route and assumed that the Universe was static. It would have been possible to have explored the assumption that the Universe was expanding, but in the absence of observational evidence this would have required a considerable leap of the imagination.

Einstein did not make this leap. He made many greater leaps than this in mathematical and theoretical physics, but he was not an astronomer, and he simply trusted the wisdom of the astronomers of his day.

Einstein assumed that, since his theory of general relativity predicted a dynamical cosmology, it must be *wrong* when applied to the Universe as a whole!

Nevertheless, Einstein was convinced that general relativity was basically sound: it was theoretically elegant and it made successful predictions of how gravity behaved in the Solar System. The question to him was, how could he fix it, change it a little so that it would work for cosmology too?

We have seen in Chapter 19 that Einstein solved this problem in a characteristically elegant manner, by finding a unique way to introduce a negative pressure into cosmology without giving up the principle of relativity. This pressure and its associated density would be independent of the observer, of position and of time.

Elegant or not, when Hubble discovered the expansion of the Universe, Einstein abandoned the constant. That made sense in his day. But today we can see that he may have been too hasty. We are not in his position, of trying to construct a static Universe. We have a worse problem: an accelerating Universe.

The cosmological constant in particle physics

The modern “rehabilitation” of the cosmological constant began, however, with completely unrelated developments in theoretical particle physics. Recall that Einstein had no physical model for the fluid that produces the negative pressure: the cosmological constant is *ad hoc*. Today particle physicists have a possible physical model. It seems that a cosmological constant may arise naturally in quantum theory.

Ironically, the story of the quantum justification for Λ starts with some other work by Einstein. Recall that we saw in Chapter 8 that Einstein had invented the concept of a photon. He had shown that light comes in discrete packets of energy, and the amount of energy depends on the wavelength of the light. He showed that this idea explained a number of experimental facts, but he did not work out a fully quantized theory of light to back it up.

In fact, it proved very difficult to find such a theory. Light is an electromagnetic wave, so physicists realized that they needed to invent a quantum theory of electromagnetism. Quantizing the atom, establishing the quantum theory that would predict correctly the spectral lines of different atoms: this was the work of Bohr, Heisenberg, and Schrödinger in the 1930s. But it took physicists another two decades to get a good quantum theory of electromagnetism, culminating in the independent work of the American physicists Richard P Feynman (mentioned in Chapter 21) and Julian Schwinger (b. 1918), and the Japanese physicist Sin-Itiro Tomonaga (1906–1979). Physicists finally had a theory that gave a precise meaning to the notion of a photon, almost 50 years after Einstein introduced it. The theory is called **quantum electrodynamics** (QED).

The reason for the long wait was that the photon presented physicists with some of the most difficult theoretical problems they had ever grappled with. We have learned enough about quantum physics in this book to be able to understand one of the difficulties. Einstein had shown that energy was quantized, that having more energy in the electromagnetic field meant having more photons. But unfortunately, there is the Heisenberg uncertainty principle. We met this principle in Chapter 7, where we saw that it forces atoms in a gas to have a certain zero-point energy, which can't be removed by cooling the gas. The wave oscillations of the electromagnetic field also have a minimum zero-point energy, which can't be removed. So although there may be *no* photons at all, quantum theory tells us that there is still some energy in the field. This purely quantum effect was something that Einstein could not have anticipated in his early work on photons, long before Heisenberg.

The zero-point energy of oscillation of atoms in a gas presents no real problems. The energy per atom is small, and there are only a finite number of atoms. But the problem gets more serious when one considers electromagnetism, the theory

In this section: one reason for expecting the cosmological constant to exist is that it comes out naturally from modern approaches to quantizing electromagnetism. In fact, the puzzle is that it should be very much larger than it is.

►The three scientists shared the 1965 Nobel Prize for physics for QED.

of light. Photons are bundles of energy associated with electromagnetic waves of a particular wavelength, and the uncertainty principle requires a zero-point energy for the vibrations of *each* different wavelength. But there are an infinite number of possible wavelengths! Even if there were no photons around at all, there would be an infinite amount of energy associated with the zero-point vibrations. One of the successes of QED was showing how to deal with this zero-point energy – how to describe the way that charged particles affect one another and give off photons when accelerated, without being disturbed by the infinite energy that seems to pervade space.

The zero-point energy is controlled but not eliminated in QED. In fact, it leads to experimentally verified predictions. One of the most striking is the so-called Casimir effect, named for the Dutch physicist Hendrik Casimir (1909–2000). Imagine an idealized experiment, where two metal plates are placed parallel to one another a small distance apart. The plates are infinite in extent in both directions, and we imagine that they are perfect conductors of electricity, which means that they are shiny mirrors that reflect all photons. The electromagnetic waves between them now cannot have arbitrary wavelengths. The photons will reflect from the plates, bouncing between them. The allowed wavelengths are only those that fit exactly between the plates: one-half wavelength, one, one-and-a-half, two, and so on, just as a violin string (or a star, as in Chapter 8) has only certain allowed wavelengths or frequencies of vibration.

Now, in the idealized experiment, imagine bringing the plates closer together – to half of their original separation. Most of the originally allowed wavelengths of photons will still fit in the new separation, but one will not: the longest wavelength of the larger separation will not fit the new one. This means that its zero-point energy is no longer present in the space between the plates. By bringing the plates together, we have reduced the total zero-point energy of the space between the plates. This means that there should be an *attraction* between the plates: if by bringing them closer we liberate energy, then Nature will want to do this. This attraction exists even if we consider a more realistic case where the plates are of finite size, as long as the plates are very close together. It can be, and has been, measured. The only way to explain it is by ascribing reality to the zero-point energy of photons that are not even there!

The problem for gravity is that this energy ought to create gravity. The energy measured by the Casimir effect is only the difference between the total energy when the plates are in one position and that in another position: the difference between two infinitely large numbers, which in this case is a finite number. But for gravity, we expect that *all* the zero-point energy should make a gravitational field. If that were the case, space would curve up dramatically. So it appears not to be there, or at least not so much of it. Is there a way to get rid of this energy? Is it really there?

Let us ask in what way this energy would create gravity. There can be no special experimenter for measuring the zero-point energy, since it is a property of empty space.

All experimenters must measure the same energy, regardless of their motion relative to one another. This means that the zero-point energy has to have exactly the same property that Einstein needed for the energy created by the cosmological constant. This zero-point energy is equivalent to a cosmological constant! The attraction between the conducting plates is due to the negative pressure associated with this energy.

Considering that the zero-point energy is potentially infinite, what limits are there on its size? The only limit that physicists generally agree on comes from the fact that it does not make sense to talk about photons with a wavelength smaller than the Planck length that we first met in Chapter 21, which is the smallest length that most physicists think can be used in theories that do not embody quantum gravity. If we add up the zero-point energies of all possible photons with wavelengths larger than this, the cosmological constant we get is huge, contributing a much larger energy density than any known matter in the Universe. In fact, since the calculation can only involve Planck's constant, the Planck length, and the speed of light, it must be a number made, like the Planck length itself, out of G , c , and \hbar . And it must have the units of density, so it should be proportional to the Planck density, given in Equation 21.12 on page 295:

$$\text{natural zero-point mass density} \propto \rho_{\text{Pl}} = c^5/hG^2 = 8 \times 10^{95} \text{ kg m}^{-3}.$$

It would be unreasonable to expect the constant of proportionality to be so small that the natural mass density would be small compared to the critical density of the Universe, so although we have made the zero-point energy finite, we still have a big problem.

One way out is to postulate that the zero-point energy is really there, but that it is cancelled to a high accuracy by a cosmological constant of the opposite sign. The remainder would be the effective, observed cosmological constant. But this does not solve the problem. It just pushes the original question into a new one: why is the cosmological constant so large, so that when it cancels the zero-point energy the difference (the *effective* cosmological constant) is so small?

So the question now facing physicists is not, does the cosmological constant exist? Rather, the question is, why is it so small? At the present time, physicists have no answer to this question.

It is ironic that what Einstein regarded as his biggest mistake might yet prove to be one of his most important contributions! And not just because the Universe may be accelerating today. Let us now look at what it may have been doing when it was just a baby.

Inflation: a concept waiting for a theory

Inflation is an idea, or a working hypothesis, about what happened in the early Universe to make it so homogeneous. It has other consequences, which are now well-supported by observation: that the Universe should be almost flat, that galaxies should have formed from initial density fluctuations of a certain size and distribution. Inflation is about what tricks the zero-point energy of particles in the early Universe might have played.

Inflation can't yet be called a physical theory, because we don't know its cause. Rather, it is a phenomenon that can occur in the very early Universe if the correct theory of high-energy physics has certain properties. The circumstantial evidence for it is strong. And inflation seems to be a feature of a large class of high-energy physics theories. Cosmological observations therefore have the possibility of guiding the development of these theories.

Here is a list of questions that inflation sets out to answer.

- Q1. How did the Universe get to be so homogeneous and isotropic on the large scale? We saw earlier in this chapter how difficult it is to understand the large-scale similarity between different regions of the Universe. Inflation offers an explanation.

In this section: inflation answers many questions about cosmology, but it is not yet grounded in any fundamental theory of physics.

- Q2. Why are there no **magnetic monopoles**? This is a serious problem for particle physics, but it is one that we have not yet come across in this book. A monopole is the magnetic analog of electric charge: it would be a purely North pole, or a purely South pole. The Universe has plenty of electrically charged particles, but apparently there are none that have a single magnetic charge. The only way we get magnetism is from moving electric charges, so that North poles are always accompanied by South poles on every magnet.

► The word *monopole* means one pole: a particle that is just a single magnetic pole. Inflation was originally invented in order to explain the absence of monopoles.

But the laws of electromagnetism permit monopoles, which would create fields like charges do but with the electric and magnetic aspects exchanged. For example, a moving magnetic monopole would create an *electric* field. The standard theories of high-energy physics not only allow monopoles: they suggest that they should have been created in abundance in the early Universe. Inflation explains why they are absent today.

- Q3. Why is the density of the Universe so nearly critical? Dark matter observations (Chapter 14) and the theory of the creation of helium in the Big Bang (Chapter 25) tell us that the density of matter is within one-third of critical. The mass density associated with the dark energy carries a further two-thirds of the critical density. Considering all the possible values that the total density could have, why is it so near to critical? Inflation, at least in its simplest form, predicts that the Universe should be almost exactly critical.
- Q4. How did galaxies form: why did sufficiently large fluctuations in density occur in such an otherwise smooth Universe?

Inflation power: the active vacuum

In this section: we explain how inflation works. It relies on a change in the quantum state associated with the vacuum and a release of energy. This has analogies with phase changes in magnets.

Inflation relies on a form of the cosmological constant that arises temporarily in the early Universe. In this section we will see how such a temporary energy field can come out of the laws of physics. In the next section we will use our understanding of the active gravitational mass to show how this field drives the Universe into a rapid expansion.

Recall our earlier discussion of the Casimir effect. There we saw that, in quantum theory, the “vacuum” is a state in which particles are absent, but which still has plenty of energy, the zero-point or uncertainty energy.

Many physicists now believe that it is possible that the laws of high-energy physics allow for there to be two or more different vacuum states, with different amounts of energy, and that the temperature of the Universe determines which vacuum it is in.

This may seem a contradiction in terms: how can there be two different ways in which particles can be absent? The difference is in the way the energy of the vacuum is determined. Two analogies may be helpful in seeing that this is possible.

The first analogy is with a waterfall. Imagine that you are boating on the Niagara River above Niagara Falls. The river moves placidly and you are so far from the falls that you can't see or hear them. You look around and feel, intuitively, that you are floating at “ground level”. You have no sense that you are really high up on a plateau. Now suppose that on the following day you go fishing a long way downstream of the waterfall. Again you cannot see or hear the waterfall, so you sit on the riverbank, at “ground level”, and you have no sense that you are lower than the level you were at the previous day.

The vacuum state in quantum theory is like the “ground level”. It is not an absolute level, but just a state in which, under suitable conditions, there are no particles:

everything is quiet and placid. If the conditions change (near Niagara, you move from one place to another; in cosmology, the Universe changes its temperature), then the nature of the vacuum can change.

Let us pursue this analogy a little further. Suppose that you went from one ground level to the other by allowing yourself to drift up to the Falls and fall over them. If you manage to survive the drop over the waterfall, you will arrive at the lower ground level, but not right away, and you will not have made a smooth journey. For the Universe, the transition from one vacuum to another was also not smooth: it resulted in a huge release of energy, which created all the particles of which we are made today. But eventually, like the Niagara River, it settled down into its present placid state.

Here is the second analogy, which is actually quite a good one, since the physics and mathematics are similar: it is the formation of a “permanent” magnet. Many minerals acquire magnetism as they cool. For example, molten lava from a volcano has no magnetism. But as it cools, its molecules find that if they line up all their spins in a consistent way, then they will have a lower total energy than if their spins are randomly oriented. When the lava is hot, the kinetic energy of vibration is much larger than this spin orientation energy, so the lava does not have any systematic orientation. But when it is cooler, the random vibrations are weaker than the spin orientation effects, and the material prefers to align its spins. When all the molecules are oriented in a consistent way, their spins combine to create effectively a small electric current, and this is what creates the magnetic field of the object. Magnets that you buy in a store are made like this.

The final direction that the spin takes is essentially random, but it can be influenced. If the mineral is in a magnetic field when it cools, then the spins tend to line up with the external field. This external field just gives them a little nudge in the right direction; it does not force them to align. They “want” to align because there is an energy benefit to do so. The alignment process releases this energy, so for a brief time the mineral is reheated slightly.

Inflation in its simplest model is very similar to the process that produces magnetism in minerals. When the Universe was very hot, the laws of physics were simple. As we mentioned before, the standard view of physicists is that all the forces of Nature were then on an equal footing, all with the same strength. Then, as the Universe expanded and cooled, a different state became the preferred one. In this sense, there was an “alignment” in the abstract space of all possible strengths of forces and masses of particles. In this picture, some details of this alignment were random. This is called **spontaneous symmetry breaking**. In this process, a large amount of energy was released. Unlike the analogy with magnetism, where the energy difference is small, in this case the energy difference was huge. It is this energy release which created the Universe as we see it today.

The energy released is the zero-point energy of the initial vacuum state. At first it behaved like a cosmological constant, with no preferred rest frame. But eventually the energy was transferred to other fields, creating the photons, neutrinos and quarks of the very early Universe. During the cosmological constant phase, the Universe expanded rapidly. This was the epoch of inflation.

The switch from one vacuum state to another is called spontaneous symmetry breaking because the simplicity (symmetry) of the original vacuum is replaced by (broken into) the complexity we see in particle physics today. This is a kind of

►This is just a thought experiment: don't try it yourself!

►The magnetization of cooling lava provided one of the crucial pieces of evidence in favor of plate tectonics, the theory that the continents move around on the Earth. The Atlantic ocean is widening at the mid-Atlantic ridge, a deep furrow running North–South roughly midway in the ocean. Geologists found that the direction of the natural magnetism of the rocks on the ocean floor near the ridge alternates between North and South as one moves away from the ridge: on either side of the ridge there are alternating bands of magnetism. The explanation lies in the periodic reversals of the Earth's magnetic field. The alternating bands imply that the rocks are formed and then move away from the ridge, new ones being formed at the ridge to replace them as the Atlantic widens.

change of phase in the early Universe. Theories that unify the strong, weak, and electromagnetic interactions are called Grand Unified Theories (GUTs), and they all have to use spontaneous symmetry breaking to explain the fact that there is no unity today among these forces. In fact, there have been several epochs in the Universe where this happened, and each may have left its mark.

This is indeed the way many physicists think the Universe evolved, although study has shown that the inflationary phase may have been entered and left more gradually than the analogy with magnetism suggests. The most common guess is that inflation happened when the temperature of the Universe was about at an equivalent energy of $kT = 10^{16}$ GeV, which is about $\times 10^{-3}$ of the Planck mass-energy. This is called the GUT energy scale. This is below the Planck energy, but not by much, so inflation would have happened very early.

Inflating the Universe

In this section: inflation expands the scale-factor of the Universe exponentially, at a very early stage in the expansion of the Universe.

Now we can describe how inflation works. The dynamics of the inflationary Universe are remarkable.

To see what to expect, look at Equation 24.12 on page 362, but replace the density ρ with the active gravitational mass $\rho + 3p/c^2$, which gives the general expression for the acceleration of the expansion. This gives

$$a_{\text{cosmol}} = - \left(\frac{4\pi G}{3} (\rho + 3p/c^2) \right) d. \quad (27.1)$$

When the dominant forms of energy and pressure are given by the vacuum energy, which has the same properties as the cosmological constant, then $\rho + 3p/c^2$ is negative, equalling -2ρ . While the negative pressure exerts no local forces because it is uniform, it actually causes the Universe to expand. This would not happen if gravity were governed by Newton's law of gravity, where only the mass density creates gravity. But Einstein's theory allows inflation, and that is the crucial difference.

The expansion produced by this vacuum energy is particularly rapid. In Equation 27.1 we can do some simple dimensional analysis to get the time-scale. The left-hand side is an acceleration, which has dimensions of distance divided by the square of time, and so the right-hand side must have the same dimensions. The right-hand side contains the distance d , so the remaining factors must together have the dimensions of 1/time². Thus, a characteristic time in the problem is obtained by taking the inverse square-root:

$$\tau = \left(\frac{3}{8\pi G|\rho + 3p/c^2|} \right)^{1/2} = \left(\frac{3}{8\pi G\rho_v} \right)^{1/2},$$

where ρ_v is the vacuum energy. This is the characteristic time-scale for the expansion. The expansion is exponential on this time-scale, as we noted in Chapter 25.

What was the density at the GUT scale when inflation may have happened? The density when the temperature was equivalent to the Planck mass-energy was presumably the Planck density, Equation 21.12 on page 295. At this time the Universe would have been radiation-dominated (rest-masses were probably unimportant in the energy density), so the density decreased as the fourth power of the temperature. Since the temperature went down by a factor of 1000, the density decreased by 10^{12} to a mere 10^{84} kg m⁻³! With this density, the time-scale for exponential growth evaluates to about 10^{-38} s. The Universe roughly doubles its size every time the clock ticks 10^{-38} s! The full equation for the exponential time-dependence of the scale-factor is

$$R(t) \propto e^{t/\tau}.$$

A region the size of the Planck length at the beginning of inflation would reach a macroscopic size, say 1 mm, in only 73τ . So inflation does not need to last long to make a huge difference.

This exponential expansion can't go on forever, because the energy being released is converted into normal matter, whatever that is at the GUT scale! It presumably behaves like radiation, with a positive pressure. Without a big negative pressure, the exponential expansion ceases, and the Universe starts to decelerate. But it does so from the enormous initial expansion speed provided by inflation. This is the point where the standard Big Bang picture begins to take over. At a time later than 10^{-38} s but probably earlier than 10^{-30} s, the Universe is a hot gas of normal matter (the old vacuum energy) in a state of the present vacuum.

Inflation put to the test

Although inflation is not yet grounded in a theory of fundamental physics, scientists have actively explored various “scenarios”, sets of assumptions about how inflation might behave in detail. These serve to restrict the possibilities for fundamental theories by eliminating variants that do not fit observed data.

The earliest full version of inflation, proposed by the American physicist Alan Guth (b. 1947), turned out to be too simple: it produced too much density irregularity today by ending too quickly. Subsequent work has focused on “slow-roll” inflation, which ends more gradually and gives acceptable agreement with the density irregularities needed to explain galaxy clustering. One interesting variant, called chaotic inflation, works even if the initial conditions before inflation were highly variable from one place to another. In regions where the Universe was initially contracting, inflation never took place, and so human beings were never created. In this picture, we happen to be part of a patch that was initially expanding. This is relevant to our discussion of the Anthropic Principle, below.

In an earlier section we wrote down a list of problems that inflation tries to solve. Now that we know what inflation does, we can see how it produces solutions.

- A1. It is easy to see how inflation solves the homogeneity/isotropy problem. If the period between the onset of inflation and its cessation is long enough, the expansion would have inflated any small region into an enormous size. The Universe we see today could have come from something very small, so small that even at the early time of 10^{-38} s it would have had time to smooth itself out in the relatively quiescent period before inflation began. In this picture, the distant galaxies and the various regions of the Universe at decoupling were all part of the same original tiny domain.
- A2. This also shows how inflation solves the monopole problem. The reason that inflation is assumed to have occurred around the GUT energy is that the Universe reached this after forming monopoles. So even if monopoles were abundant before inflation, they will be dispersed so far apart that the chance of our encountering one now would be minimal.
- A3. Inflation also solves the problem of why the Universe is so close to its critical density. The reason is in the conditions at the end of inflation. The exponential expansion phase wiped out any memory of the initial expansion velocity before inflation set in. In the exponential expansion, the Hubble parameter is just the reciprocal of the time-scale τ . Its square is then

$$H^2 = \frac{1}{\tau^2} = \frac{8\pi G \rho_v}{3c^2}.$$

In this section: observations confirm most of the predictions of inflation.

By Equation 26.1 on page 388, the universe has zero spatial curvature: it is flat! Of course, real inflation can only be approximately exponential, and it has to make a transition to ordinary expansion, so this equation will only be a first approximation. But it implies that after the Universe exits from the inflationary phase, it must remain nearly flat, with a density close to the critical density.

Notice that the inflation-dominated universe has a very special geometry. Because the cosmological constant fluid has the same density and pressure to all experimenters at all times, regardless of their motion, it follows that this universe model must look the same to all observers: unlike our present Universe, there is no preferred rest frame. This symmetry implies that the geometry of space is actually flat, with the galaxies flying apart from one another through it. The curvature of time in this model is produced entirely by the Doppler redshift of these galaxies relative to each other.

►The inflation-dominated universe model was discovered by the Dutch mathematician and astronomer Willem de Sitter (1872–1934) in 1917, immediately after Einstein introduced the cosmological constant. Nevertheless, de Sitter disliked the cosmological constant and argued (long before Hubble's observations) that general relativity implied that the Universe was expanding.

- A4. Finally, inflation provides an initial spectrum of density irregularities at an early time that can lead to galaxy formation. It does this by amplifying initial quantum fluctuations in the quantum fields that describe matter in the very early Universe. Although such fluctuations are initially tiny, they increase in size during the period of inflation.

Just at the beginning of inflation, the Universe is unstable: a small random fluctuation in density can initiate inflation in one place before another. Since inflation then changes the density exponentially with time, the density contrast between two places that start inflating at slightly different times gets larger and larger, amplifying by the cube of the factor by which the Universe expands. In this way a tiny quantum fluctuation can grow to the size needed to begin galaxy formation.

The details of how this density fluctuation now produces galaxies are very sensitive to assumptions one makes about the transition from inflation to the normal expansion. It also relies on the dark matter in the Universe, since this is free to start collapsing as a result of this overdensity, while the ordinary protons and electrons are tied to the photons. But numerical simulations give excellent agreement with observations so far, as in Figure 25.3 on page 380.

Is inflation still going on?

In this section: the acceleration of the Universe is evidence that some kind of weak inflation is happening today.

We have seen in Chapter 25 that observations of Type Ia supernovae have suggested that the Universe is accelerating even today, although at a much smaller rate than during the epoch of inflation. Detailed studies of the cosmic microwave background irregularities have independently given further evidence for this. This is shown in Figure 27.2.

While both inflation and the acceleration we see today are similar to the behavior of a universe model with a positive cosmological constant, it seems likely that something more complicated than a cosmological constant is driving this acceleration. Inflation, certainly, was not caused by a cosmological constant, simply because it was not constant: the epoch of inflation came to an end.

Scientists are therefore looking for a theory of a variable dark energy field, which can act for limited times, changing its strength in a natural way. As we noted in Chapter 19, some physicists call their proposals *quintessence*. The search for such a theory is in its infancy, and it may require much more data, both from physics experiments and from cosmological observation, before believable models can be found. But as long as astronomers continue to believe that the Universe had one or more periods of accelerating expansion, physicists will continue looking for an

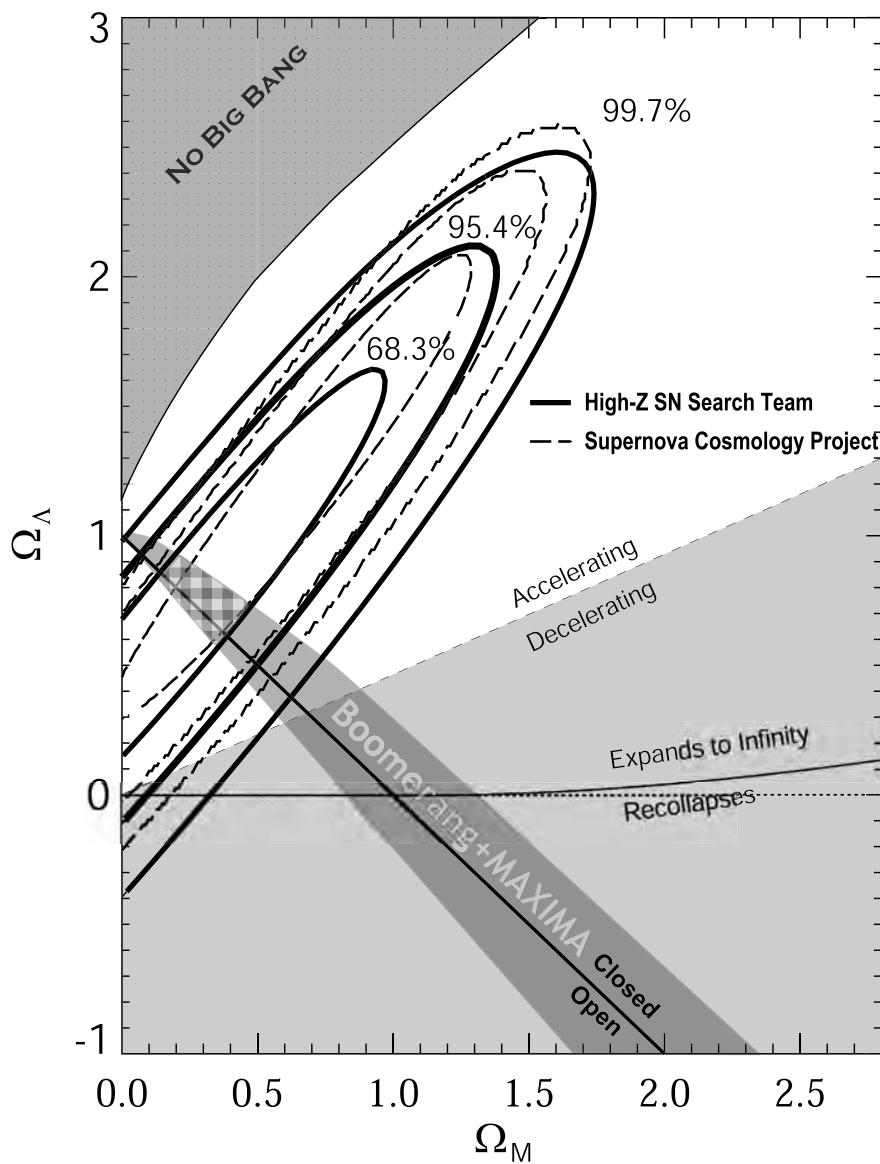


Figure 27.2. This chart shows the implications for the large-scale nature of the Universe of the measurements of Type Ia supernovae and of the sizes of the fluctuations in the microwave background, according to the data available in 2001. The horizontal axis is the fraction of the critical density that is in matter, $\Omega_m = \rho/\rho_c$. The vertical axis is the fraction in dark energy, Ω_Λ , that is contributed by the cosmological constant Λ . If these add to one, then the Universe is flat: its total mass density is equal to the critical density. The downward sloping solid line is the line where these two numbers add to one. The region above the line has more mass, and so represents closed universes; the region below represents open universes. The various oval regions show the parts of the diagram consistent with the observations of Type Ia supernovae, as in Figure 24.4 on page 352. (The percentages drawn in the figure are confidence levels on the observational uncertainties: the observations tell us that the true Universe must lie within the outer ovals with a probability of 99.7%.) The dark wedge around the closed/open line is the region allowed by the observations of the microwave fluctuations. There is only a very small part of the diagram where the two observational constraints overlap, and this is shown as the hatched area. The center of this area is a model where the mass density of the Universe is 30% of the critical density and the density contributed by the cosmological constant is 70% of the critical density: a flat expanding Universe. Figure from the High-Z Supernova Search, based on data from de Bernardis, P., et al (2000) *Nature* **404**, 955, and Balbi, A., et al (2000), *Astrophysical Journal* **545**, 1.

explanation. The acceleration could turn out to be the key clue pointing toward the next theory of fundamental physics.

Is Einstein's law of gravity simply wrong?

In this section: instead of dark matter and dark energy, maybe general relativity is wrong. This is possible, but an alternative theory is hard to find.

In trying to explain the many mysteries of the Universe, we have assumed that Einstein's gravity provides a good description of both the local dynamics of galaxies and the global Universe. From this it follows that puzzling observations – such as the high velocities of matter in galaxy clusters or the strong bending of light in gravitational lenses – tell us that there is hidden mass and dark energy.

However, general relativity is simply a theory of physics, and it must always be tested against observation. In the Solar System and in the Hulse–Taylor pulsar, it passes tests superbly well. It explains black holes and neutron stars. But it is hard to test the theory over the long distance-scales of galaxies and galaxy clusters. It is not surprising, therefore, that some scientists have posed an interesting question: could the missing-mass problem be solved essentially by changing the law of gravitation over long distances, without introducing a hidden form of matter? Is there really hidden mass, or just hidden gravity?

This is a natural question to ask, but it has not led to much progress in providing viable alternatives to the model of inflation with cold dark matter that the large majority of astrophysicists favor. There are two reasons for this. One is that it seems exceedingly difficult to modify general relativity on long length-scales without throwing out its successful predictions. The other reason is that the accumulating evidence for missing mass puts strong constraints on new theories.

This last point may seem surprising, but it comes about because any modification of the laws of physics must be *universal*: it must be the same everywhere in the Universe. If there is a new term in the law of gravitation, containing (say) a new fundamental constant of Nature that determines how strong it is and over what distance-scale it is going to be noticeable, then this must give a consistent explanation of the dynamics of every galaxy and galaxy cluster in which there is a missing-mass problem. There are dozens of well-studied clusters and galaxies, and there is sufficient variety among them to challenge any of the proposed modifications of gravity.

In particular, there seems to be no single length-scale on which missing gravity takes over from Newtonian or Einsteinian gravity. This is not surprising if one accepts that the missing gravity is created by missing mass: each galaxy or cluster condensed around an individual clump of dark matter, and since these clumps are random, the gravity they create will be different from galaxy to galaxy and cluster to cluster.

One kind of theory that does not suffer from this problem has gained much attention among physicists in the last few years. These are called “brane-world” theories, and they are inspired by string theory. We will look at them later in this chapter, and in particular we will see that some of them explain the dark matter as ordinary matter existing in other three-dimensional worlds separated from ours by a small distance in a fourth spatial dimension. In these theories gravity can bridge the gap between these worlds but the other forces of physics cannot, so we are unaware of their existence except for their gravitational influence on us.

The book is certainly not closed on new theories. If the dark matter particle is detected, much of the motivation to look at ideas like these will disappear. But if dark matter searches show that the required particles are not there, then scientists will take these theories much more seriously.

Cosmic defects

Inflation is only the most spectacular example of what can happen if modern theories of particle physics are applied to early cosmology. Whether or not inflation occurred, we might still find the Universe filled with what are called “defects”, primarily cosmic strings and cosmic textures. These also arise from spontaneous symmetry breaking at lower energies.

When a symmetry breaks spontaneously, it breaks randomly. In different parts of the Universe, the values of certain fields in the theory will have random aspects that differ from one another from place to place. In some kinds of theories, these differences lead to cosmic strings, which are long thin condensations of trapped energy. Inside the string, space is still in the old “false” vacuum, which has lots of energy relative to the present “true” vacuum. If the string arose from symmetry breaking at an energy of 10^{16} GeV, then the energy can be enormous: the strings can have a mass per unit length of 10^{21} kg m⁻¹.

However, their gravitational behavior is not simply that of a massive piece of rope. Since the matter inside is trapped vacuum energy, its energy density ρ is accompanied by a negative pressure $p = -\rho c^2$. In this case, since the string is one-dimensional, the pressure acts in only one direction, and provides a tension along its length that keeps the string together. But the active gravitational mass is $\rho + 3\langle p \rangle/c^2$, where $\langle p \rangle$ is the average pressure. Since the pressure is zero in the two directions perpendicular to the string, the active gravitational mass is just $\rho + p/c^2$, and this vanishes!

So a cosmic string does not curve time. Clocks are not redshifted by it and particles do not go in orbit around it, despite its immense mass.

How, then, can a string have any effect, and in particular how can it assist the formation of galaxies?

The primary effect of the string is to curve space. In Chapter 19 we defined the active curvature mass, $\rho - p/c^2$, but this was valid only for isotropic pressure. For strings, the result is more complex, and needs to be calculated from the details of Einstein’s theory. The result is that the curvature in the direction along the string is zero: there is no curvature mass, and so proper distance along the string is the same whether measured inside or outside the string. On the other hand, curvature in a plane perpendicular to the string is non-zero, and it is generated by a density of curvature mass equal to 2ρ .

This has observational consequences. For example, we have seen that light deflection by the Sun depends on the spatial curvature as well as the curvature of time. Therefore, a cosmic string will deflect light that passes by it. This would lead to a kind of gravitational lensing in which one might get double images of a star, one from light that passes one side of the string and the other from light passing the other side. Astronomers have looked for this effect, but so far without success.

Our simple picture of the geometry around a string gets more complicated if two strings intersect, or if a string becomes dynamical and has oscillations, as it would be expected to do. Strings can break off when they intersect, and form closed loops. Dynamical strings give off gravitational waves (by shaking the curvature in the plane perpendicular to their length), and loops that do that will shrink and eventually disappear. Moving strings are also good seeds for galaxy formation: although time is not curved, the curvature of space causes the geodesics of particles near the string to deflect toward the loop. When this happens, densities increase in the string’s wake, particles collide, and galaxies can form.

In fact, scientists have calculated that there is a simple relation between the

In this section: cosmic strings could have been formed in phase transitions in the early Universe, and could produce detectable effects today.

amount of gravitational radiation that strings emit and their effectiveness as seeds for galaxy formation. Current limits on the amount of radiation are placing constraints on the number of strings, but they are not yet able to eliminate the theory that cosmic strings formed galaxies.

Observations of the microwave background are, however, placing much tighter constraints on cosmic strings. At short angular scales, strings produce a very different pattern of fluctuations than inflation. At this time (2002) it seems unlikely that strings were plentiful enough to have been the main cause of galaxy formation. But they could nevertheless still be there in large quantities, and could be observable through their gravitational effects, including gravitational radiation.

Cosmic rays

In this section: the mystery of the highest-energy cosmic rays has the potential to introduce new physics into astronomy.

Cosmic rays are charged particles, mainly protons, that strike the Earth at very high energies. They are detected by looking for their collisions with atoms of the atmosphere. Historically, cosmic rays were the way physicists first studied what we now call "high-energy physics", before they could create high-speed particles in laboratory accelerators. The muon, one of the three types of leptons (see Chapter 25), was first discovered in the products of the collisions of cosmic rays with atoms of the atmosphere, and the time dilation of special relativity was convincingly verified by observing that fast-moving muons in cosmic rays lived much longer than ones at rest (Chapter 16).

Today cosmic rays are once again pushing physicists to the limits of what they can understand. The highest-energy cosmic rays are observed with energies above 10^{20} eV, which is far beyond any energy that can be produced in an accelerator. But what is challenging about them is not the physics of their collisions with other particles. What is puzzling is that there are so many of them. As we noted above, the Earth encounters roughly one such particle every second.

These high-energy particles do not seem to come from the Galaxy, because their arrival directions do not coincide with the plane of the Milky Way, and they are too energetic to have been deflected into their arrival directions by the small magnetic fields in interstellar space. On the other hand, they should not be able to move too far between galaxies, because they would lose energy to interactions with the cosmic microwave background radiation. When a low-energy photon of the microwave background collides with such a high-energy particle, many things can happen, including the production of other particles. This would cause a high-energy proton to lose its energy rapidly.

Calculations suggest that high-energy protons or other known elementary particles should not be able to move further than about 20 Mpc before their energy falls below about 10^{20} eV. Yet, within this distance (about the distance from the Earth to the Virgo Cluster), and in the directions from which the various particles have been seen to arrive, there are no visible sources. One would expect that any astrophysical object that could produce protons of such extraordinarily high energy would be doing something else as well, like producing light or X-rays or gamma-rays. But astronomers see no likely candidates.

The arrival directions of these particles are not known to high precision, so there are certainly galaxies from which they could have come. But the galaxies do not look like they contain anything special. The speed of these particles is so close to the speed of light that we would expect to see the event in which they were produced almost at the same time as we receive the particles. No supernovae or other spectacular events have been associated with observations of these particles.



Figure 27.3. One detector of the Pierre Auger Observatory, an array of 1600 similar units being constructed in Argentina, able to observe high-energy cosmic rays striking anywhere in an area of 3000 km^2 . It is hoped that this will provide enough data to solve the mystery of the ultrahigh-energy cosmic rays. Image courtesy Pierre Auger Observatory.

And certainly there is no evidence for such events happening once every second!

Other events, much further away, would be more likely to produce particles at these energies. Gamma-ray bursts, for example, or perhaps quasars. But these do not happen so close to our Galaxy.

There is as yet no explanation of these high-energy cosmic rays. New instruments are under construction that will gather much more data and hopefully lead to a solution. The solution might be simple, such as a form of particle acceleration that scientists have overlooked and which is present in all galaxies. The particles might be very heavy nuclei, which would have the observed energy at a speed slow enough to avoid rapid energy loss. But it is also possible that we are being presented here with some completely new physics. Perhaps the sources are relatively nearby but dark (cosmic strings, magnetic monopoles, decaying massive dark matter particles, ...), or perhaps there is new physics in the cosmic-ray particles themselves. The resolution of this problem certainly has the potential to affect the other issues we are discussing in this chapter.

Quantum gravity: the end of general relativity

We have arrived now at the limits of general relativity. Almost everything in the earlier chapters has been standard physics. Even though black holes may seem exotic, they are well-understood theoretically and they have been identified observationally. They are not particularly controversial in physics today. Gravitational waves have not yet been detected directly, but there is little theoretical doubt about their existence and general properties. Our ignorance about the large-scale structure of the Universe is still high but the framework provided by Einstein's equations seems adequate to describe the evolution of the Universe that we see today.

We have also seen some more speculative physical ideas, especially in the first part of this chapter: inflation, the cosmological constant, cosmic strings. These are not so well-established, and some of them might either fall out of fashion tomorrow, or be turned into "standard physics" by a crucial astronomical observation next year. But all of these ideas are rooted in fundamental physics as we now understand it. The negative pressures and cosmic defects of these speculations are features that are expected from theories that describe the nuclear interactions. The speculative part is whether the correct theory will turn out to exhibit these features at just the right energy (or temperature) to explain the facts we observe.

Big as these speculations are, there is an even bigger hole in physicists' theories.

The biggest incompleteness in physics has to do with gravity. Just as gravity drives the evolution of the Universe and of most things in it, gravity also drives the most fundamental and exciting theoretical research in physics today. Gravity is where the action is, if you are a fundamental theoretical physicist.

The reason is that Einstein's general relativity is not, cannot be, the last word on gravity. General relativity is what physicists call a classical theory of physics. It has none of the distinctive features of quantum theory, and that is a contradiction that must be fatal for general relativity.

The reason is simply the uncertainty principle. Consider the fact that in quantum systems, one cannot measure exactly how much energy the system has, or exactly where it is located. Nevertheless, gravitation theory tells us how to compute the gravitational field from the distribution of energy. So we could in principle measure the gravitational field far away with arbitrary precision (if it is a non-quantum field) and determine what the distribution of energy in the spacetime is with arbitrary precision, contradicting quantum theory.

In this section: the goal of theoretical physics is to unify gravity with the other forces and produce a quantum theory of gravitation. There are many situations where a quantum theory is needed.

The only way out of this contradiction seems to be to invent a theory of gravity that is a quantum theory, but which in appropriate circumstances makes predictions so close to those of general relativity that they also satisfy the observational evidence that supports general relativity. This is a standard situation when one wants a quantum theory of phenomena that are already well-described by a classical theory. The theory of electromagnetism does a good job on electric and magnetic fields that are used in everyday life, such as those that create and are created by electrical circuits in the home. But when one wants to describe electromagnetism on the level of single photons interacting with single electrons, then one needs the quantum version of the theory, QED.

In the same way, scientists need a quantum theory of gravity in order to study some phenomena with confidence. Here is a partial list of the places where quantum effects in gravity might make important changes from the predictions of general relativity.

1. *Singularities* (Chapter 21). The center of a black hole contains a place where, according to general relativity, time finishes. Any object that falls in and reaches this location will not progress further in time. Most scientists find this disturbing, and hope that the strong gravitational fields near a singularity will create substantial uncertainties in a quantum theory of gravity, and that these uncertainties will allow particles near the singularity to continue into the future indefinitely. However, it is also possible that a quantum theory of gravity will not remove singularities but embrace them in some way, so that they are no longer places where physical theory breaks down.
2. *Hawking radiation and naked singularities* (Chapter 21). There seems to be little doubt in the minds of most physicists working on these questions that black holes must emit radiation with a spectrum basically like that of a black body, but of course this cannot last forever. When all the mass of the hole has been radiated away, something must happen to the singularity inside. Does it become “naked”, i.e. visible to the outside world? Does it disappear altogether? Or does it remain hidden behind a horizon that has shrunk to a point? In general, one might expect quantum gravity to say something about the cosmic censorship hypothesis, which we discussed in Chapter 21.
3. *The Big Bang* (Chapter 24). This was a singularity in our distant past. Perhaps a quantum theory will tell us how the Big Bang came about, whether there is any meaning to the notion of time *before* the Big Bang, whether there were many Big Bangs with different outcomes, whether the Big Bang was smooth or bumpy.
4. *Planck scales* (Chapter 21). The characteristic numbers that we call the Planck mass, length, time, and so on, are built from Planck’s constant (quantum theory), the speed of light (relativity) and G (gravity). A quantum theory may make very different predictions about physics on these scales, which could only be attained in rare circumstances. In particular, on distance-scales shorter than the Planck length, spacetime might not even be continuous. Some physicists suggest that spacetime really has the structure of a tangle of disjoint loops, and only looks smooth when averaged over distances larger than a Planck length. Others suggest that it consists of tiny fluctuating wormholes (see below and Chapter 21), called spacetime foam. Still others think that spacetime might really have ten or eleven dimensions, the extra ones (above

the four of conventional general relativity) being visible to us only over distances of the order of the Planck length. (See the discussion of branes below.)

5. *Negative energy, wormholes, and time travel* (Chapter 21). We have seen that to keep wormholes open for travel one needs negative energy. It may be that quantum gravity can supply a source of such negative energy. If so, it is most likely to occur only on the Planck scale, leading to the idea of spacetime foam. But it is possible that quantum gravity will teach us how to make sustainable regions containing negative energy. The applications of this would be immense, and would come close to elements of the science fiction world of space travel. In particular, it would theoretically be possible to travel through a wormhole and emerge earlier, having traveled backwards in time.
6. *Shadow matter*. This is a long shot, but one that could cause enormous difficulties with gravitational experiments. The possibility exists that a theory that unifies all the forces of Nature will predict a class of matter that interacts with the rest of matter only gravitationally. Within this class there could be a complex series of interactions between its particles: they could have their own charges, strong forces, and weak interactions, but these would be insensitive to the ones that we know. In effect there could be a hidden Universe occupying the same space as we occupy, with its own structures and dynamics. It would be detectable through its gravitational effects, but would otherwise be dark and invisible. The missing mass could in principle come from this sector of the theory, and if dark matter experiments fail to detect weakly interacting particles then this may become a real possibility. But experimental confirmation would be exceedingly difficult to provide.
7. *Branes*. String theory, the leading candidate today for a unified theory of all the forces, and therefore the leading candidate for providing us with a quantized theory of gravity, requires that the real spacetime that we occupy have eleven dimensions, and that our four-dimensional world is just a subspace, a kind of membrane, in the larger space. This higher dimensionality is not arbitrary; the theory can only be made mathematically self-consistent in this number of dimensions. The way that our four dimensions fit into the larger spacetime is not known, but it is possible that they are like a twisted, kinked ribbon, called a **brane**. According to these models, the non-gravitational forces of physics act only in the brane but gravity can exert an influence over a region of the eleven-dimensional space near our brane. The initial assumption that this region would be as small as a Planck length has recently been challenged by some physicists, who suggest that it could be much larger, up to say 1 mm. If true, this so-called brane-world picture could have observational consequences: modifications of Newton's law of gravity at very small distances, generation of gravitational radiation in the very early Universe, creation of gravitational waves and gravitational attraction by matter that does not exist in our Universe but rather inhabits another nearby brane (referred to earlier in this chapter), explanation of the nature of the early Universe without invoking inflation, and perhaps more. It is interesting and exciting that even rather simple gravitational experiments, such as measuring the force of gravity over short distances, could in principle provide evidence for quantum gravity.
8. *Gravitons and the definition of energy*. It is tempting to expect that a quantized theory of gravity should involve quantized gravitational waves, which

▷ In fact, physicists working in string theory are excited about its prospects because it is the only theory that they have been able to make self-consistent. There is a feeling among some that there may be only one possible mathematical structure that can be made self-consistent, and it must therefore be the correct theory.

physicists call gravitons. However, the concept has problems. The usual idea would be that they should be like photons, which carry quantized amounts of energy that depend on the photon's wavelength. The problem is that energy is not so easy to define in general relativity, and indeed even gravitational waves go away if one goes to a locally inertial frame and looks on a scale smaller than the wavelength of the wave. Gravitons need to be a little like Lewis Carroll's Cheshire Cat: if you look too closely they go away! This illustrates an aspect of quantum gravity that is not often discussed: how quantizing gravity will change conventional *quantum theory*. It seems likely that quantum gravity will inherit from classical general relativity its inability to define energy in an invariant way. Far from giving us a concrete picture of the graviton, quantum gravity might instead make our picture of the photon a little fuzzier!

In this list I have tried to anticipate how quantum gravity might affect predictions about phenomena that we can already study in classical general relativity. But quantum gravity need not, surely will not, confine itself to modifying what we already know. In the rest of this chapter I will speculate on areas where quantum gravity might solve entirely new problems or introduce entirely new ideas.

A Universe for life: the Anthropic Principle

In this section: we examine the fine-tuning of physical quantities that seems to have been required for the evolution of life. Many of these arguments can be found in two stimulating books, *The Anthropic Cosmological Principle* by J D Barrow & F J Tipler (Oxford University Press 1986) and *The Life of the Cosmos* by L Smolin (Oxford University Press 1997).

We have already seen in Chapter 11 that the appearance of life on at least one planet in the Universe requires some special values of some fundamental physical quantities. If these values were substantially different, life could not have evolved. Quantum gravity, by unifying all the forces of Nature, could be in a position to explain how some or all of these values came to be. There might be no arbitrary parameters, no adjustable values. Quantities like the mass of the proton, for example, should either be predicted by the theory or be given a certain probability, from which we shall in principle be able to calculate what the probability was of having values that would have led to the evolution of life.

Here we expand the list of these numbers beyond those we mentioned in Chapter 11. We will take the view that the values of Planck's constant h , Newton's constant G , and the speed of light c are not numbers that have to be predicted by the theory. Their values depend on the human-based system of units in which we express them. In fact, we have seen in Chapter 21 that they just define a natural system units in which everything else can be measured. So we ask about how the evolution of life depended on certain dimensionless numbers, like the ratio of the mass of the proton to the Planck mass, which play a role in the physics that we have discussed in this book. We start with the ratio of the mass of the electron to that of the proton.

1. *Ratio of the mass of the electron to that of the proton.* If the mass of the electron were much larger, say comparable to the mass of the proton, then the structure of atoms would be very different. Electrons would orbit very close to the nucleus, and the energy required to ionize an atom by removing the electrons would be much larger. Chemistry would be totally different. While life might still be possible, it would be on terms that we would not recognize.

On an astronomical level, the structure of white dwarfs would be rather different, since the electrons would contribute considerably to their self-gravity. This would lower the Chandrasekhar mass (see Chapter 12) and perhaps make it impossible for supernova explosions to occur, since (as we also saw in Chapter 12) the process that produces the explosion is finely balanced, and a much lighter neutron star might not release enough energy to drive the outer lay-

ers of the giant star away. This would reduce the production of the heavy elements needed for life.

2. *Ratio of the mass of the proton to the Planck mass.* If the mass of the proton were larger, then again this would lower the Chandrasekhar mass, and supernovae might not occur. Even a few percent increase in the proton mass might have this effect, stifling the evolution of life.

On the other hand, if the proton mass were smaller, the Chandrasekhar mass would be bigger, and the collapsing core might be too massive to stop at the neutron star stage: its self-gravity might overwhelm the nuclear forces and lead to the formation of a black hole directly. This would again deprive the collapse event of the strong rebound shock wave that blows the envelope away.

The evolution of life has been very sensitive to having exactly the right ratio of the proton mass to the Planck mass.

3. *Difference between the proton mass and the neutron mass.* The neutron has slightly more mass than the proton. This difference is large enough to allow a free neutron to decay into a proton, an electron, and an anti-neutrino. This decay is called beta decay. But in a nucleus, neutrons can be stable against beta decay: it takes a little more energy to make a proton in a nucleus than outside it, because of the presence of other nearby protons whose electrostatic repulsion of the new proton raises its total energy. This is not a large energy, so the existence of stable nuclei depends on the neutron having a mass very close to that of the proton. If it were much larger, all nuclei would decay by beta decay, and life would be impossible.

If the neutron had a mass smaller than the proton by even a small amount, then electrons and protons would spontaneously combine to form neutrons, releasing the extra energy as a neutrino. Life would again be impossible, because there would be no chemistry: the Universe would consist only of collections of neutrons bound together by the nuclear attraction.

Like the proton, the neutron has only a small mass range available to it in which life can evolve.

4. *Energy levels of the carbon nucleus.* We discussed this in detail in Chapter 11: the synthesis of elements heavier than helium depends sensitively on the details of the energy levels of the carbon nucleus. A small change in these levels, caused by a small change in the strength of the nuclear force, would make life impossible.
5. *Strength of the hard-core nuclear repulsion.* Another way in which the nuclear forces affect the possibility of life is that the strength of the repulsion that takes place when two nucleons (protons or neutrons) come close to one another has an effect on the formation of neutron stars. If the repulsion is not strong enough, then when a white dwarf core of a giant star collapses (having the Chandrasekhar mass, which is determined only by the proton mass, not by the nuclear forces), the collapse will not stop before a black hole is formed. Recall that the neutron star has a radius that is only about three times larger than its Schwarzschild radius, so it only has to collapse to one-third of its radius in order to form a black hole. A modest decrease in the strength of the hard-core repulsion could allow this.

On the other hand, if the hard-core repulsion were to increase, then neutron stars would form with much larger radii and lower densities. The binding energy released would be smaller, and this could have the consequence (as we saw above) that the envelope of the giant star would not be blown off. In both cases, supernovae might not occur, some heavy elements would not form, and life as we know it would not be possible.

Therefore – as with the proton mass, the neutron–proton mass difference, and the longer-range nuclear forces – there is only a small range of values of the strength of the nuclear hard-core repulsion which will allow life to evolve.

6. *The mass of the electron neutrino.* We have seen in Chapter 11 that there is strong evidence that the mass of the electron neutrino is not zero, but it is still very small. If it were much larger, then the neutron would be unable to decay, since it would not have enough mass to produce the masses of an electron, a proton, and an anti-neutrino (whose mass will be the same as that of the electron neutrino). This would distort the process of forming nuclei, since there would be no beta decay. The abundances of some of the elements that are used in the chemistry of life would be much less, and the chemistry of life would be very different, if not impossible.
7. *The mass of the dark matter particles.* As of this writing (2002), we do not yet know the nature of the dark matter. But if we assume that galaxies formed because of the dark matter, then a crucial requirement for the formation of life is that the dark matter particles had a mass heavy enough that they would be cold long before decoupling. The reason is that *galaxies themselves are essential for life*. The collapse of clouds of gas onto the dark matter condensations led to heating of the gas and the formation of stars. And it was crucial that stars formed in galaxies, not in isolation. Even if stars somehow had formed without galaxies, so that they were randomly sprinkled through the expanding Universe, then there would have been only one generation of stars. When they died, the matter they expelled through winds and explosions would have simply swirled through the spaces between stars, but would never have attained enough density to form another generation of stars. So the elements made by the first generation of stars would not have found their way into new stars, planets, and people. Galaxies therefore played a vital role in the evolution of life, encouraging the first generation of star formation and then retaining the gas released by those stars and mixing it into the giant molecular clouds from which second, third and subsequent generations of stars formed. These stars had the heavier elements that the first generation did not have, and they could form planets. The Earth is made almost entirely of atoms that were created in stars and held in place by the Galaxy until the Sun could form, and with it the planets of the Solar System. If the mass of the dark matter particles were too small, galaxies would not have formed at all.
8. *The violation of time-reversibility in fundamental physics.* We have seen in Chapter 25 that the violation of time-reversibility in the fundamental laws of physics led to the fact that the Universe had slightly more protons than anti-protons, slightly more neutrons than anti-neutrons. Since we are made of this excess, its existence is crucial to the evolution of life. If the Universe had been completely symmetrical between matter and anti-matter, then almost all the particles would have annihilated against one another, and the only particles

left in the Universe would be a few lucky ones that had avoided encounters with anti-particles. This particle–anti-particle mixture would have been fairly uniform, so that any clumping of the kind that led to stars in our Universe would just have led to further annihilations as particles and their anti-particles got closer together. Stars could not have formed in such circumstances, and of course neither could life.

It is sobering to reflect on the fact that this subtle and almost unobservable violation of time-reversal invariance, which is a feature of physics that even physics students generally do not learn much about until they reach graduate school, is one of the foundations of life itself! We can hope that quantum gravity will explain the violation of time-reversal invariance.

9. *Balance of dark matter and dark energy.* The Universe left the inflationary epoch with the critical mass density, but inflation does not prescribe how that density is shared between dark matter and dark energy. If there had been much less dark matter, then galaxy formation would not have occurred, and, as described above, life would not have been possible.

Biologists and chemists debate whether the evolution of life was inevitable, given the original conditions on the Earth. Astronomers try to estimate how many Earth-like planets the Galaxy might contain. These are arguments about probabilities, about how many places in the Universe might contain life like ours. However, if any of the physical parameters in the above list were substantially different, then probabilities would have been irrelevant: life as we understand it would simply have been impossible.

It may be that quantum gravity will tell us that these parameters were inevitable, and so there is no need for further discussion. But this seems to me to be unlikely. Instead, quantum gravity could predict the probability that the Universe had begun with the values of these parameters. Then we will be able to assign a probability that the Universe could have evolved life anywhere. This probability may come out rather small. If that happens, what will that mean?

As we mentioned in Chapter 11, the Anthropic Principle is relevant here. The existence of human beings implies that we live in a Universe in which the evolution of life was possible, so our measurements could not have come out any other way. But this is unsatisfying if there is one and only one Universe. If, however, quantum gravity tells us that there have been many universes, each beginning with random values of the constants, and the number of such cosmic experiments was unlimited, then the problem goes away: there will inevitably be universes in which the constants take the appropriate values, and we live in one of them.

Some physicists have speculated on how this might work out. Maybe the Universe is really bound and will re-collapse to a Big Crunch, and by quantum effects re-expand with another Big Bang with different fundamental constants. This seems unlikely to be a good explanation, since one of the fundamental constants is the expansion speed: if the speed of one such re-expansion is so large that the Universe never re-collapses, then the process comes to an end, and maybe not soon enough to have produced life with high probability. In fact, the process seems to be ending with our present Universe, which apparently will not re-collapse. It seems too much of a coincidence that the cyclic Universe would stop cycling just when life evolved on a tiny planet in an unremarkable corner of an anonymous Galaxy.

Another possibility is raised by inflation. If the Universe before inflation was very inhomogeneous, then different regions could have inflated in different ways. If quantum gravity tells us that these different regions had different values of the fundamental constants, then our problem might go away. We inhabit just one of many regions of the entire Universe, and in our region the parameters allowed life to evolve. In a neighboring region, the parameters might be very different. This region is far away now, too far for us to see. While this explanation could be correct, it would not be verifiable, and that is unsatisfactory.

Whatever the ultimate explanation turns out to be, the most important point of all is that explanations seem possible, that quantizing gravity could have much bigger implications than just increasing our understanding of gravity.

Causality in quantum gravity: we are all quantized

In this section: a quantum theory of gravity will have to address issues of time and causality beyond present quantum theory.

Another major conceptual change that quantum gravity is likely to bring about is to our notions of predictability in physics, of cause and effect. These were already modified by standard quantum physics, which removed the older, classical idea of Newton's era that one could at least in principle predict arbitrarily accurately what the outcome of any experiment would be, provided one had sufficiently good information about the physical conditions at the starting point. In quantum theory, one can only hope to predict a set of *probabilities* about the outcome of an experiment. If one does the experiment very many times with identical starting points, then one can test the prediction, verifying that the frequency with which any outcome appears is consistent with the probability predicted by quantum theory. In quantum gravity, even this amount of predictability may be eroded.

One source of trouble is that conventional quantum theory can assign probabilities to outcomes of the experiment only by making a sharp distinction between the experimenter and the experimental system; the experimenter remains classical and behaves with free will, setting up the system for the experiment as many times as desired, measuring the outcomes. But quantum gravity seems unlikely to be able to do this. For one thing, it must presumably give us a quantum cosmology, that is a quantum theory of everything, in which there is *no* outside observer to do experiments and to predict the frequency with which something happens. We are part of the cosmology! If quantum gravity can give us a logically consistent way of dealing with such a situation, then when it is applied to a laboratory experiment it is likely to merge experimenter and experiment into one, and remove the complete freedom that the conventional experimenter has to perform the same experiment over and over again.

By telling us how to merge the experimenter and the experiment into one system, quantum gravity may also help us to understand some existing experiments that are consistent with conventional quantum theory but which seem inconsistent with normal notions of causality. It is possible to create so-called entangled states, where for example two particles are created in such a way that their total spin is zero but the spin of each one is not fixed. Then measuring the spin of one determines the spin of the other. It is possible to show in these experiments that the spin of a particle is not determined at the time it is created and then carried along with it, in the way a classical particle would behave. The spin is not determined until it is measured. Up until that point, there are only probabilities of one spin value or another. However, once the first particle has been measured, the experiments show that measuring the spin of the second particle always gives the right value, the opposite spin to the first particle so that the two spins add to zero. So it appears that, by measuring one particle, one puts the other into a definite spin state too. But when the particles are so far apart by the time they are measured that it is possible

to measure the second before any information has had time to travel from the first to “tell” it what spin state to assume, the second particle is still in the correct spin state! How can this information have traveled from one particle to the other? If it did not, how did physics conspire to give a correlated result? Knowing how to treat experimenters and their experiments as a single relativistic system might help resolve this paradoxical behavior, and quantum gravity ought to tell us just that.

Another puzzle for causality would be time travel through wormholes. It would be difficult to sustain a view that a person traveling back in time had the freedom to set up experiments or do anything else that we would call “free will”, because then one falls into the “grandmother paradox”: go back in time and murder your grandmother when she is still a little girl. Do you exist? Paradoxes like these seem to require that, if time travel is possible, the objects that travel back in time behave consistently in the past; quantum gravity must choose one kind of behavior in all of spacetime, and nothing is free to vary that. The grandmother paradox is no paradox because, just as you are about to run over your juvenile grandmother with a car, your car’s tire bursts, you hit a tree, and you are killed! Your attempted assassination becomes one of your grandmother’s favorite stories, one you heard when you were small, but which you did not realize was going to be about you! You have no free will, and the quantum state of the Universe is what it is and always has been: your time travel is just part of it.

A final source of difficulty for causality might be the evaporation of black holes. In conventional quantum theory, no information is in principle lost when a system evolves. The outcomes are only predicted with some probability, but if we do the experiment many times we will find that all aspects of the initial state will have some influence on the outcome. But with black holes this appears not to be the case. If we start with a star that is about to collapse to a black hole, there are some aspects of the distribution of mass inside the star that do not show up in the final state, which is an evaporating black hole. Much information has just been swallowed by the hole. When the hole evaporates away entirely, we will again have a smooth spacetime, but one that bears no information about some aspects of our starting spacetime. Black holes seem to be information destroyers of a fundamental kind. It may be that quantum gravity will rescue this situation, and will predict that the information we lack is actually hidden in tiny correlations among the photons emitted as the hole evaporates, or that the final spacetime is not as smooth as we have assumed. Or quantum gravity may just force us to live with a further erosion of our ability to predict things.

The quantization of time?

Playing with causality may seem like disturbance enough to our way of thinking about the world, but quantum gravity will go beyond this: it will play with *time* itself. In a way, time has been the main sub-theme of this book. Gravity expresses itself mainly through the changes it makes to time. The gravitational slowing down of time (the gravitational redshift), the curvature of time that is central to Einstein’s explanation of how gravity works, the fact that time itself comes to an end for a particle that encounters a singularity, and the related fact that time itself began at the Big Bang – all of these are part of the fundamental connection between gravity and time.

Time is, however, among the most puzzling of concepts in modern physics. What *is* time? Why does it have its one-way character? There are many so-called *arrows of time*, ways of recognizing that time moves in one direction but not the other:

- *the psychological direction*, the fact that we have memories of the past but

In this section: time is one of the great mysteries of physics. A quantum theory of gravity must shed light on this, since time is just part of our geometry.

not of the future;

- *the statistical direction*, which means that heat always flows from a hotter body to a colder, that different gases can easily mix together but never spontaneously separate;
- *the wave direction*, the familiar observation that waves of any kind always move outwards from their source, never converging inwards on the system, as would happen in a film of a wave run backwards in time;
- *the quantum measurement direction*, where a measurement wipes out previous uncertainties in a quantum system, so that (for example) a particle can no longer reach a location after a measurement that it had a non-zero probability of reaching before it;
- *the cosmological direction*, the expansion of the Universe that appears to have the same direction everywhere, and which may not ever reverse; and
- *the fundamental-physics direction*, the tiny violation of time-reversibility in the laws of physics that is intimately connected with the fact that the early Universe had more matter than anti-matter.

Physicists and philosophers debate the relationships between these arrows of time, questioning whether only one is fundamental, and the others derivable from it, or alternatively whether there is a coincidence of two or more independent arrows that point in the same direction.

It is to be expected that quantum gravity will alter our notion of gravity and therefore of time. Perhaps it will allow us to give a meaning to “before” the Big Bang and to “after” a singularity. Perhaps it will illuminate the relationships between the different arrows of time. Perhaps quantum fluctuations in gravity will lead to quantum fluctuations in the sense of time. Most intriguingly, perhaps quantum gravity will give us a more radical notion of time, as an average over something more fundamental that happens on sub-Planck-scale dimensions.

We have seen that the Planck length is the smallest length-scale on which it is sensible to think of space in our conventional way, as a continuous background within which things happen. What is physics like on sub-Planck scales? Guesses and speculations abound. It could be a kind of spacetime foam, composed of Planck-mass black holes that fluctuate into and out of existence on Planck timescales. It could be a tangled web of strings or loops with no particular sense of dimension or direction. It could even be a set of discrete points linked together by mathematical relations that make them look continuous on larger scales. And if space has a messy structure on small scales, so too must time.

Time for the twenty-first century

In this section: we stand on the threshold of a revolution in gravitation theory as big as Einstein's. Most exciting of all is what the new theories may do to our concept of time.

The idea that time is not continuous allows physicists to begin to ask questions that would have been unthinkable in Einstein's time. How does time emerge from the tangled mess at sub-Planck scales? Why is there just one time dimension, out of many space dimensions? Is time a kind of *mistake*, a defect in a mathematical structure that normally would produce only space directions on scales larger than the Planck scale? Will quantum gravity teach us how to control time, to manipulate its direction or “flow”? Or, as hinted in the previous section, will quantum gravity simply take away our freedom to stand outside a physical system and manipulate it, even as it gives us the knowledge of what we could do with time if only we could do so?

In this book we have journeyed through the Universe, starting from the Earth and ending in the contemplation of the Universe on its largest and smallest scales. We have also made a journey through the world of scientific ideas. We started in Galileo's world, gaining deep and profound insights from simple experiments with cannonballs and pendula. We coupled his and Newton's insights into gravity with the understanding of atomic physics that scientists developed during the century from roughly 1850 to 1950, opening up the physics of planets, stars and galaxies. Then, halfway through our journey, we entered Einstein's conceptual universe, introduced at the beginning of the twentieth century. New ways of thinking about space, time, and matter have helped us appreciate the richness of the physical Universe revealed by the incredible blossoming of astronomy that took place in the second half of the twentieth century.

I write this just as we are beginning the twenty-first century, and the blossoming of astronomy shows no sign of abating. With new and more powerful instruments, astronomers are certain to continue to surprise us with new discoveries, at both astronomy's interface with particle physics and astronomy's interface with biology. At conferences, lectures, and press conferences, the excitement that astronomers feel for their subject is palpable: they are on the trail of some of the deepest secrets of the Universe, and they are getting new clues to these secrets every day.

Just as exciting is the promise the new century brings of a genuine revolution in our thinking: quantum gravity. This promise drives the work of thousands of the world's most talented theoretical and experimental physicists. When a good theory of quantum gravity finally arrives, it may or may not lead to technologically useful by-products, tools for society like microchips or nuclear energy. But it will surely create a new universe of ideas, one that may require as big a change in our thinking as the universe that Einstein created. It will re-define our understanding of cause and effect. It will illuminate the earliest steps on the long road that led to the evolution of our own lives.

Most profoundly – dwarfing all the other developments we may foresee in the realms of gravity, astronomy, cosmology – quantum gravity must re-define time itself. Time is at the heart of gravity. It slows and curves to make the planets orbit the Sun, it stops entirely inside a black hole, it doesn't even begin to advance for a photon or a gravitational wave. And for intervals shorter than the Planck time, it may not behave like time at all.

To my thinking, the most profound result of quantizing gravity, the most important reason for encouraging – and joining! – the efforts of the thousands of physicists and astronomers who are trying to solve the puzzle of how to unite the visions of Albert Einstein and Max Planck, will be to discover how the *quantum nature of gravity* leads us to an understanding of the *quantum nature of time*.

Appendix:

values of useful constants

The following values are useful in the exercises and in evaluating equations in the text. All values are quoted in SI units, which are reduced to the fundamental units. Thus, the units (dimensions) of G are given as $\text{m}^3 \text{s}^{-2} \text{kg}^{-1}$ rather than the equivalent $\text{N m}^2 \text{kg}^{-2}$. The only exception is the Hubble constant, which is given in its conventional units.

Symbol	Value	Units	Description
c	2.9979×10^8	m s^{-1}	speed of light
G	6.6726×10^{-11}	$\text{m}^3 \text{s}^{-2} \text{kg}^{-1}$	constant of gravitation
g	9.807	m s^{-2}	acceleration of gravity on Earth
m_e	9.1094×10^{-31}	kg	mass of the electron
m_p	1.6726×10^{-27}	kg	mass of the proton
m_n	1.6749×10^{-27}	kg	mass of the neutron
h	6.6261×10^{-34}	$\text{kg m}^2 \text{s}^{-1}$	Planck's constant
k	1.3807×10^{-23}	$\text{kg m}^2 \text{s}^{-2} \text{K}^{-1}$	Boltzmann's constant
eV	1.6022×10^{-19}	$\text{kg m}^2 \text{s}^{-2}$	electron volt
r_e	2.8179×10^{-15}	m	classical electron radius
σ	5.6705×10^{-8}	$\text{kg s}^{-3} \text{K}^{-4}$	Stefan–Boltzmann constant
M_\odot	1.989×10^{30}	kg	mass of the Sun
R_\odot	6.9599×10^8	m	radius of the Sun
L_\odot	3.826×10^{26}	$\text{kg m}^2 \text{s}^{-3}$	luminosity of the Sun
R_\oplus	6.3782×10^6	m	equatorial radius of the Earth
M_\oplus	5.976×10^{24}	kg	mass of the Earth
y	3.1557×10^7	s	length of a year
AU	1.4960×10^{11}	m	radius of Earth's orbit about the Sun
p_{atm}	1.013×10^5	$\text{kg m}^{-1} \text{s}^{-2}$	atmospheric pressure on the Earth
pc	3.0857×10^{16}	m	parsec
H_0	70	$\text{km s}^{-1} \text{Mpc}^{-1}$	Hubble constant
H_0^{-1}	4.4×10^{17}	s	Hubble time

The SI units include a number of standard prefixes that indicate how a unit is changed by a power of ten. The table below lists the standard prefixes and their symbols (e.g. k for "kilo").

Size	Prefix	Symb	Size	Prefix	Symb	Size	Prefix	Symb
10^{-24}	yocto	y	10^{-3}	milli	m	10^9	giga	G
10^{-21}	zepto	z	10^{-2}	centi	c	10^{12}	tera	T
10^{-18}	atto	a	10^{-1}	deci	d	10^{15}	petra	P
10^{-15}	femto	f	10^1	deka	da	10^{18}	exa	E
10^{-12}	pico	p	10^2	hecto	h	10^{21}	zetta	Z
10^{-9}	nano	n	10^3	kilo	k	10^{24}	yotta	Y
10^{-6}	micro	μ	10^6	mega	M			

Astronomers prefer to use their unit of a parsec (pc) rather than call it 30 exameters. In fact, astronomers are non-conformists in many ways. They still quote measurements in the older CGS system (centimeter-gram-seconds) rather than in SI units, so that astronomy publications are full of references to centimeters, grams, and ergs. In this book we stay with the SI units, except that we also use parsecs as a distance measure.

Glossary

Terms in the glossary are printed in **bold** where they first appear in the text.

absolute magnitude A measure of the intrinsic luminosity L of a star, defined as $M = -2.5 \log(L/2.9 \times 10^{28} \text{ W})$. Because of the minus sign, stars with lower values of M are intrinsically brighter. A star that is five magnitudes brighter than another is 100 times brighter. *See also* apparent magnitude.

absorption spectrum The spectrum (color content) of the light from a star typically has dips in intensity at certain wavelengths, where elements in its outer atmosphere absorb the light preferentially. The details of the absorption are a fingerprint of the chemical composition of the outer part of the star and also carry information about its temperature, pressure, and density.

acceleration of gravity Near the surface of the Earth, this is the acceleration with which all bodies would fall to the ground if there were no resistance from air or other forces. Typically called g , it has the value 9.8 m s^{-2} . The acceleration at other distances from the center of the Earth or near other bodies will have different values.

accretion The process whereby gas falls onto an astronomical body. This can be via an accretion disk confined to a plane or in a more spherical way. The term is not used for the assembly of large amounts of gas to form a body in the first place; it is used when the body subsequently acquires smaller amounts of material. *See also* accretion disk.

accretion disk When gas falls from one star onto another in a binary system, it typically spirals around the second star before reaching it. The spiralling material forms a disk. If the second star is very compact, then the gas in the disk can reach high temperatures, where it will emit X-rays. Disks can also form around planets and individual stars during their formation. *See also* accretion.

action at a distance The term that describes the fact that the force of gravity in Newton's theory of gravity acts between separated bodies without any intermediary and without any delay. This is different from electromagnetism or general relativity, where waves of the field go from one body to another.

active curvature mass The term used in this book for the combination of density and pressure that is the source of the curvature of space in general relativity: density - pressure/ c^2 . *See also* active gravitational mass.

active gravitational mass The term used generally for the combination of density and pressure that is the source of the curvature of time in general relativity, which is responsible for most ordinary gravitational effects: density + 3pressure/ c^2 . *See also* active curvature mass.

angle of inclination Used to describe the orientation of an orbit from the perspective of a viewer on the Earth. It is defined as the angle between the line-of-sight and a line perpendicular to the plane of the orbit. A system with an inclination of 90° is one that is seen edge-on to the orbit.

Angstrom A commonly used but non-si unit of distance, defined as 10^{-10} m . This is typical of the sizes of individual atoms.

angular momentum The quantity that measures the amount of spin a body has. It is defined with reference to a particular point. In Newtonian physics and in special relativity, the angular momentum of a single particle of mass m moving on a circle of radius r about the reference point with speed v is the product mvr . For larger bodies the angular momentum is the sum of the angular momenta of all the particles in the body. When the motion is not circular, only the component of the velocity perpendicular to the radial direction from the reference point is used. In general relativity, the angular momentum can only be defined in certain circumstances, particularly when the space-time is invariant under rotations about the reference point. *See also* momentum, component, conservation of angular momentum, invariance.

anisotropy The property that a system is not the same in all directions from a given point; the opposite of isotropic. *See also* isotropic, homogeneous, inhomogeneous.

Anthropic Principle Really a collection of several principles, variants on the theme that the Universe contains human beings because it was designed to contain them. The strong version assumes that the Universe was created with this intent; the weak version of the principle merely says that scientists cannot expect to observe a Universe that could not have created people, so that certain observed conditions are inevitable.

anti-gravity The term that describes a situation in which gravity is repulsive rather than attractive. This does not happen in Newton's theory, but can happen in general relativity when the pressure is so large and negative that the active gravitational mass is negative. This underlies the theory of inflation in the early Universe. *See also* inflation, cosmological constant, dark energy, quintessence.

anti-matter All elementary particles have a counterpart, called an anti-particle, that has the same (positive) mass, the same spin, and opposite electric charge; when a particle collides with one of its anti-particles, they annihilate each other and convert their mass-energy into photons or other particles. Anti-matter is the collection of all anti-particles of normal matter. The Universe began with an excess of matter over anti-matter;

otherwise the particles of which we are made would have been annihilated long ago. *See also* anti-proton, positron, photon.

anti-proton The anti-particle of a proton. *See also* anti-matter.

aphelion The point on an elliptical orbit around the Sun that is furthest from the Sun. If the central object is a star, the point is called the *apastron*; if the Earth, *apogee*. *See also* perihelion.

Apollo Name given to the US program begun in the 1960s to send men to the Moon. It consisted of a number of missions, at first in Earth orbit, then orbiting the Moon without landing, and finally a succession of landings. The first landing was Apollo 11 in 1969. The Apollo 13 mission nearly ended in disaster, but the astronauts successfully returned to Earth. The earlier Apollo 8 mission caught fire on the launch pad, killing its astronauts. The scientific legacy of the program is important. The astronauts returned with rock samples that have shown that the Moon was formed from the debris from an enormous collision between the Earth and a body the size of Mars. The astronauts left behind reflectors on the Moon that are still used to track the Moon's orbit to an extraordinary precision.

apparent magnitude A measure of the apparent brightness (energy flux) F of a star, defined as $m = -2.5 \log(F/2.4 \times 10^{-8} \text{ W m}^{-2})$. Because of the minus sign, stars with lower values of m appear brighter. A star that is five magnitudes brighter than another is 100 times brighter. *See also* absolute magnitude.

arrow of time The term that describes the perception that time advances in the same "direction", never reversing. The psychological perception of time is probably related to one or more different arrows of time that physicists have identified. These include the increase of entropy (disorder) with time, the spreading of radiation and waves outwards with time, and a tiny lack of time-symmetry in the fundamental laws of physics. *See also* entropy.

asteroids Small rocky bodies in the Solar System. They may be residues of a planet that did not quite form. Most orbit the Sun on roughly circular orbits, but sometimes encounters with one another place an asteroid on an orbit that plunges toward the Sun. The collision of one such body with the Earth about 60 million years ago is thought to have been responsible for the extinction of the dinosaurs. *See also* Kuiper Belt.

astronomical unit The mean distance of the Earth from the Sun, about 1.4960×10^{11} m.

atoms Basic units of matter, which combine to form the chemicals of which our world is made. Atoms consist of a small but massive positively charged nucleus composed of protons and neutrons, and a much lighter and larger cloud of electrons. *See also* proton, electron, nucleus, chemical elements, isotope.

bar detector A gravitational wave detector made of a metal cylinder, which is stretched into longitudinal oscillation by a gravitational wave. *See also* gravitational wave, interferometer.

baryon A collective name for protons, neutrons, and related unstable particles of larger mass. The total number of baryons

(with anti-particles counted negatively) is conserved in nuclear reactions. Electrons are leptons, not baryons. *See also* lepton.

beaming A term used in special relativity to describe the effect of Lorentz–Fitzgerald contraction and time dilation on the direction of radiation from an accelerated charge. The faster the charge goes, the more its radiation is directed in the forward direction. *See also* Lorentz–Fitzgerald contraction, time dilation.

beta decay A form of radioactivity which is driven by leptons. Normally a particle decays to produce, among other particles, electrons or positrons, plus associated neutrinos. Because leptons interact with one another by weak interactions, beta decay usually happens on a longer time-scale than radioactivity involving rearrangements of baryons, such as nuclear fission. *See also* baryon, lepton, neutrino, weak interaction.

Big Bang The name given to the beginning of the Universe, which seems to have occurred in a single explosion. The term was coined by F Hoyle.

Big Crunch The name given to the hypothetical end of the Universe, should it re-collapse to an infinite density. The evidence today is that it will not re-collapse, but instead progress to the Big Freeze. *See also* Big Bang, Big Freeze.

Big Freeze The name given to the hypothetical end of the Universe, should it continue to expand forever. The evidence is that this is likely to happen. *See also* Big Bang, Big Crunch.

black bodies A technical term for bodies that are perfect absorbers of radiation. Such bodies emit a characteristic spectrum of radiation that depends only on their temperature and not on their composition. *See also* black-body radiation.

black-body radiation The radiation emitted by a perfect black body; its spectrum depends only on the temperature, and its intensity on the surface area. Stars emit radiation that is a good approximation to the black-body spectrum, and black holes emit the Hawking radiation, which has a black-body spectrum. The larger the temperature, the shorter the typical wavelength of the emitted radiation. *See also* black bodies, Hawking radiation.

black holes Bodies that have such strong gravity that light cannot escape if it is emitted from within a certain region, whose boundary is called the horizon. Since nothing travels faster than light, black holes trap everything that gets within the horizon. *See also* horizon.

blueshift The shortening of the wavelength of radiation, which can be caused by motion or by gravity. *See also* redshift.

bolometric magnitude The brightness of a star, measured in magnitudes, using the light in a range of colors defined to span the visible spectrum. *See also* apparent magnitude, absolute magnitude.

bore waves A shock in water waves, which can build up into a high wall of water moving upstream. Seen on several rivers with strong tidal ranges.

boson Each elementary particle has spin, an intrinsic angular momentum that, because of quantum effects, always is either an integer or half-integer multiple of $h/2\pi$, where h is Planck's constant. Particles that have integer spin are called bosons, those with half-integer are fermions. Bosons have a preference for occupying the same quantum state, so that they bunch together. Photons are bosons, and laser light is an example of the way they try to conform to one another. *See also* fermion, quantum theory, photon, laser.

boson star A hypothetical star composed of a different and hypothetical form of matter: a boson field. If the mass and (repulsive) self-force of the field have suitable values, then it is possible to make stars of a mass and size similar to neutron stars. No such particles are known from experiment, but grand unified theories allow them. *See also* neutron star, boson, grand unified theories.

brane A technical term in string theory, which is the current leading contender for the way of unifying gravity with the other forces of Nature. In string theory, elementary particles are not point-like, but are instead represented by small closed loops (strings) in a space with ten dimensions. Subsurfaces of this space with more than one dimension generalize the notion of strings, and are called branes, from the word membrane. *See also* string theory.

bremsstrahlung The electromagnetic radiation emitted by a rapidly moving charged particle when it is suddenly decelerated. If the initial speed is large and the deceleration great, the radiation is strongly beamed in the forward direction. *See also* beaming.

brown dwarfs Astronomical objects intermediate in mass between large planets and small stars. By definition, they do not have enough mass to raise their interior temperature to the point where nuclear reactions ignite; instead, they glow by radiating away their gravitational potential energy. Not many are known, but it is possible that there is a huge population of them that contributes substantially to the mass of our Galaxy.

Brownian motion The random movements of, for example, a speck of dust floating on the surface of water. Collisions with water molecules impart tiny changes in the speck's motion, which individually are invisible, but which happen so frequently that they randomly accumulate into apparently sharp changes in the motion of the speck. The speck executes a "random walk" across the water surface.

buoyancy The force that acts upwards on a body that is immersed in a medium of greater mean density, such as a hot-air balloon.

C-field A hypothetical field postulated by Hoyle and colleagues, which would be required in order for the Universe to obey the postulates of the Steady-State model of the Universe. *See also* Steady-State model of the Universe.

calculus The mathematical theory that deals with rates of change of functions. Invented by Newton and independently

by Leibniz, it provides systematic ways to solve for the motions of bodies acted upon by forces, but goes well beyond this in being able to treat variations in anything, such as surfaces (curvature), areas, and much more. Calculus is the fundamental mathematical tool of physics: all the basic laws of physics are expressed fundamentally in the language of calculus.

cataclysmic variable A class of variable star in which there are large outbursts of visible light and X-rays caused by mass transfer onto a white dwarf from a giant star that is its companion in a binary system. *See also* white dwarf.

catalyst An agent that promotes a chemical reaction (or other process) without itself being changed by the end of the process. Normally the catalyst is modified during the process but is restored to its original state by the end. Most car exhausts today have catalytic converters, in which a catalyst like platinum helps to convert pollutants into harmless gases. Unless the catalyst is degraded by other chemicals, it will continue doing its job indefinitely.

caustics Places where light rays that start from the same source and that pass through a complicated optical system are made to intersect. You can easily see caustics by looking at light reflecting from a choppy water surface; the caustics are the edges of the bright regions that flicker past the eye.

celsius The standard temperature system in most of the world (apart from the USA) and in science. The zero is defined as the freezing point of water, and the boiling temperature of water is 100 C. Formerly widely known as the centigrade scale, since there are 100 degrees between freezing and boiling water. *See also* kelvin.

centrifugal effect The apparent outward force that a body experiences when executing circular motion. The circular motion is itself accelerated, and the centrifugal effect is actually caused by whatever force causes the body to move from a straight line. The centrifugal effect is an example of an "inertial force".

characteristic frequencies All material bodies vibrate when disturbed. Bodies of finite size normally vibrate freely with a set of frequencies that depend on their composition and shape. These are called the characteristic frequencies of the bodies.

chemical elements The building blocks of all the materials of our environment. The atoms corresponding to each element are identical except, possibly, for the number of neutrons in the nucleus, and their chemical behavior – the way they combine into compounds – is the same for all. *See also* atoms, isotope.

chirp A gravitational wave with increasing frequency and amplitude, emitted by a binary system whose orbit is shrinking because of the emission of gravitational radiation. As the orbit shrinks, the orbital period also goes down and the stars or black holes speed up. These effects make the gravitational radiation frequency and amplitude increase. *See also* chirp time, gravitational wave.

chirp time The time-scale on which a chirping binary system changes its frequency by a factor of two. A chirping binary is a

binary system whose orbit is shrinking because of the emission of gravitational radiation. *See also* chirp.

classical A term used by physicists to describe theories of physics that do not incorporate quantum effects. *See also* quantum theory.

co-latitude Latitude is one of the coordinates we normally use for locating places on the Earth. It runs from -90° at the South Pole to $+90^\circ$ at the North Pole. The co-latitude measures the same angle but in a different way, starting at 0° at the North Pole and finishing at 180° at the South Pole. If the latitude is called β , then the co-latitude is $\theta = 90^\circ - \beta$. When mathematicians discuss the geometry of abstract spheres, they normally use the co-latitude rather than the latitude as one of the so-called spherical coordinates.

cold dark matter If the dark matter inferred from astronomical observations consists of particles much more massive than normal atoms and carrying no electric charge, then it would have cooled off more rapidly as the Universe expanded, and could have formed massive clumps that later attracted normal matter gravitationally and began the process of galaxy formation. This cold dark matter is the standard model of galaxy formation. However, the hypothetical particles have not yet been identified or detected. *See also* dark matter, hot dark matter.

collisionless gas If a gas is so rarified that collisions between its particles are very rare, it is called collisionless. If the particle velocities are random the gas might still behave like a normal gas, with a definite pressure and temperature. This can happen if, for example, the gas is confined by walls with which the particles collide and exchange energy.

color index The difference of the blue magnitude and the visual magnitude of a star. Since these magnitudes are defined by logarithms of the brightness of the star, this difference depends on the ratio of the brightness of a star in the visual color band to that in the blue band. This ratio is independent of the distance to the star, since both brightnesses fall off with distance in the same way. It is therefore a measure of the intrinsic color of the star, and thus of its temperature. *See also* visual magnitude, color of a star.

color of a star The color of a star depends on the relative intensities of different colors in its light. Most stars look more-or-less white to the naked eye, because the black-and-white sensitivity of the eye is greater than its color sensitivity for weak light sources. But when the colors in the light are measured, stars turn out to have different balances. Some have much more red light than blue, some more blue than red. These differences reflect differences in the temperatures of stars: cooler stars are more red. *See also* color index, black-body radiation.

component Mathematicians use the word *vector* to describe a directed quantity, like velocity. The piece of the velocity along any particular direction is called the component along that direction. To describe a vector in three dimensions requires three components. The concept of a component is used also for tensors, where it refers to the elements of the matrix that rep-

resents the tensor in a particular coordinate system. *See also* vector, matrix, tensor.

compose In the context of this book, scientists use this term to describe how different velocities combine in special relativity. When body A measures the velocity of B and B measures that of C, the velocity of C as measured by A will not be simply the vector sum of the two previous velocities. If this were the case it would be easy to get velocities greater than the speed of light. Instead, relativity predicts a more complicated composition law, which never produces a speed exceeding that of light.

Compton scattering The scattering of photons from charged particles. Since photons carry electric fields, they interact with electric charges. The scattering, however, reveals the discrete nature of photons: they behave just like particles carrying a given energy and momentum and scatter from the charged particle into various directions. *See also* photon.

conservation of angular momentum In Newtonian physics and in special relativity the total angular momentum of any isolated system is constant in time. The parts of the system can exchange angular momentum, but the total is unchanged. In general relativity, this law holds only if the geometry of the spacetime is invariant under rotations about the reference point for the computation of the angular momentum. *See also* angular momentum, invariance, conservation of energy.

conservation of energy In Newtonian physics and in special relativity the total energy of any isolated system is constant in time. The parts of the system can exchange energy, but the total is unchanged. Each time physicists have uncovered a new force, a new branch of physics, they have found that there is an associated energy that can be defined in such a way that the total remains conserved. This is not arbitrary: it is only possible to define conserved energy if the geometry of spacetime is time-independent. In general relativity, this law can therefore only hold in certain circumstances: particles moving in time-independent geometries, or in the locally flat geometry of special relativity sufficiently near to any event, or in terms of total energy as measured by a distant experimenter sitting in the flat spacetime far away from an isolated star or black hole. *See also* energy, experimenter, invariance, locally flat.

convection When a fluid is heated from below too rapidly for ordinary conduction or for any radiative flux through the fluid to carry it away, then the fluid begins to flow in a roughly circular motion, absorbing heat at the bottom of the convection cell and releasing it at the top.

Copernican principle Copernicus argued that the Sun was at the center of the Solar System, not the Earth. This made the Earth an ordinary planetary body, not located in any special place in the Solar System. When this principle is extended to the Galaxy and the Universe, we would assume (unless there is evidence to the contrary) that our location is similarly not privileged. In particular, the Universe should look the same, statistically, to any other astronomer observing it from any other ordinary star in any other ordinary galaxy. This principle is sometimes called the *principle of mediocrity*.

cosmic censorship hypothesis Solutions of Einstein's equations for black holes contain singularities inside, locations where the predictive power of the laws of physics break down. These are regarded as a mild failure of general relativity, since the unpredictability is hidden from our view behind the black-hole horizon. However, if a singularity were to form outside a black hole, this would be a much more serious problem for physics. No robust examples of this are known, however, and Penrose suggested that perhaps there was a deep connection between singularities and horizons in general relativity. His cosmic censorship hypothesis is the proposal that there should exist a mathematical theorem to the effect that in generic situations singularities never appear outside a horizon: they are always "censored" by Nature. The conjecture is unresolved. *See also* singularity, naked singularities.

cosmic microwave background radiation The early Universe consisted of a dense, hot, expanding gas. When the gas cooled, it became transparent to radiation, and most photons released at that time have traveled freely through the Universe ever since. Astronomers detect this radiation as a black-body spectrum with a temperature such that the radiation is predominantly in the microwave part of the spectrum. Tiny irregularities in temperature from one direction to another are a snapshot of a very early phase of galaxy formation. *See also* decoupling, photon, microwave.

cosmic rays The Earth is bombarded by high-energy particles, mainly protons, from space. Most collide with gas in the upper atmosphere, so the radiation poses only a limited radiation risk at sea level. Most cosmic rays probably originate in supernova explosions, but there is a small flux of ultrahigh energy particles whose origin is a puzzle. *See also* supernova.

cosmic strings Certain theories of high-energy physics predict that, when spontaneous symmetry breaking occurs to form the laws of physics as we know them, there may be some locations where the original, symmetric form of the laws still holds. These regions must be concentrations of energy that was trapped because it could not decay. If these regions are one-dimensional, they are called cosmic strings. In principle such strings could be so plentiful and massive that they caused galaxies to form near them. However, the evidence today is against this mode of galaxy formation and in favor of cold dark matter. Nevertheless, lighter cosmic string are still possible and could be sources of gravitational radiation. *See also* spontaneous symmetry breaking, cold dark matter.

cosmological constant The term introduced by Einstein into his equations of general relativity, in order to insure that there could be solutions for static cosmologies, because astronomers at that time had not found evidence for an expanding Universe. The term created a repulsive force, a form of anti-gravity, that countered the attraction of normal matter. When the expansion of the Universe was discovered, Einstein rejected this term, but today astronomers have put it back in because they have found that the expansion of the Universe appears to be accelerating. Physicists have created alternative explanations of such effects, such as quintessence. The generic name for such effects in the

equations of general relativity is dark energy. *See also* anti-gravity, dark energy, inflation, quintessence.

cosmological scale-factor The mathematical quantity that tracks the expansion of the Universe. It can be defined to be any length at any particular initial time, but then it expands in proportion to the distances between galaxy clusters that are so far apart that their mutual gravitational attraction has a negligible effect on their velocities. The expansion speed and acceleration/deceleration of the Universe are defined by how this length-scale behaves with time relative to its initial value. *See also* cosmology, galaxy cluster.

cosmology The study of the Universe as a whole, its history, and the physical processes that led generally to the formation of galaxies and stars. *See also* physical cosmology, cosmological scale-factor.

critical density The mass density that the Universe would have if the gravitational attraction of the matter was just what would be required to reduce the expansion speed of the Universe to zero in the infinite future. When there is dark energy as well, the dynamics of the Universe will be more complicated, but the ratio of the total density of mass-energy to the critical value determines the overall spatial structure of the Universe: open (ratio smaller than one), closed (larger than one), or flat (equal to one). *See also* physical cosmology, dark energy.

crust The outer layer of a neutron star, where the density is not great enough for all the matter to exist as neutrons. The crust is composed of neutron-rich nuclei in equilibrium with free neutrons and electrons. The nuclei are thought to arrange themselves in a weak lattice, which resembles a jelly-like solid. The crust is not brittle, but rather pliant and yielding. As the only likely solid part of the star, it is responsible for many observed phenomena, and could be a source of gravitational radiation. *See also* neutron star.

curvature The property of a space that determines whether parallel lines can remain parallel when extended in as straight a manner as possible. A space has zero curvature (i.e. is flat) if parallel lines remain parallel. If they approach one another the space has positive curvature, if they diverge then it is negative. Einstein used the curvature of spacetime to describe the action of gravity in general relativity. The locally straight lines are the paths that free particles follow through spacetime. The curvature of time represents, to a first approximation, Newtonian-like gravity. *See also* spacetime, spacetime-interval, spacetime metric, flat space.

cycles Ancient astronomers tried to describe the apparent motion of the planets in the sky by complicated motions superposed. If the planets are assumed to go around the Earth, then they do not always go in the same direction. Sometimes they turn around and go backwards, an effect which is easily explained in the Copernican model of the Solar System when one takes into account the motion of both the Earth and the planet. But ancient astronomers, thinking the Earth was fixed, described the motion of a planet as a basic *cycle* around the Earth, added to a smaller *epicycle* that was a kind of circular

motion back and forth along the orbit, sometimes having the net effect of moving the planet backwards.

dark energy Observations suggest that the expansion of the Universe is presently accelerating, and theories of inflation also require a period of rapidly accelerating expansion. To produce the acceleration one needs a physical field that has a negative pressure, large enough to make the active gravitational mass negative and produce an anti-gravity effect. Fields that have positive energy but large negative pressure are called dark energy. *See also* cosmological scale-factor, active gravitational mass, anti-gravity.

dark matter Astronomers have determined that the gravity that seems to bind together galaxies and galaxy clusters, and the overall gravitational field of the Universe, can only be explained if there is hidden matter, called dark matter. It is not known what it is composed of. *See also* cold dark matter.

deceleration parameter A dimensionless number that represents the acceleration or deceleration of the Universe today.

decoupling At first the Universe was hot enough to ionize hydrogen, so that its gas was a plasma of charged particles, through which photons could not move far without scattering. As the Universe cooled off, its gas became neutral, and photons could propagate. The transition is called the *decoupling* of photons from matter. It is sometimes called *recombination*. *See also* plasma, cosmic microwave background, photon.

degenerate gas If fermions are cooled and compressed to a sufficient density, then because no two identical fermions can occupy the same quantum state, not all of them can slow down to small velocities, as would classical particles. So they retain a residual pressure even as their temperature goes to zero kelvin. This is a degenerate gas, and the pressure is enough to hold up a neutron star against gravity. *See also* fermion, classical, quantum theory, kelvin, neutron star.

density parameter The ratio of the density of the Universe (or of one component of the Universe) to the critical density. *See also* critical density.

density wave In the theory of galactic structure, it is believed that a compression wave in the density of the stars and gas in a spiral galaxy is responsible for the observed spiral pattern. Most of the mass in such a galaxy is distributed symmetrically around the center, but the brightest stars are not: they are concentrated in the spiral arms. Since the time it takes a density wave to travel once around the galaxy is typically long compared to the lifetime of massive bright stars, it is believed that the spiral arms trace the location of the compression region of the wave. *See also* spiral galaxy.

derivatives In calculus, the functions that represent the rates of change of other functions. The velocity is the derivative of the position of a particle with respect to time. *See also* calculus, differential equations.

deuterium The nucleus of the isotope of hydrogen that has one proton and one neutron. *See also* isotope, nucleus, proton.

differential equations Equations consisting of functions and their derivatives. All the basic laws of fundamental physics are expressed as differential equations. *See also* calculus, derivatives.

dimensional analysis The technique of examining the consistency of an equation with the dimensions (basic units of mass, length, and time) of the quantities in it. It can be a powerful technique if, on physical grounds, the equation must contain only a few quantities with known dimensions. Then dimensional analysis can actually point the way to inventing the correct equation that relates them with one another. *See also* dimensions.

dimensionless number A number that has no dimensions (units). Arguments of non-linear mathematical functions, like the sine or exponential, must be dimensionless even if they are composed of quantities that individually have dimensions. *See also* dimensions.

dimensions The type of units carried by a physical quantity. For example, distances have dimensions of length. A distance can be given in units of meters, miles, microns, and so on, and its value will depend on the unit. But in each case, the quantity has the dimension of length.

direct image In optics, an image created by a system of mirrors and lenses which preserves the sense of left and right in the original object.

displacement A technical term for the vector position of an object from the origin of the coordinates. *See also* vector.

diverging lens A lens that causes initially parallel light rays to diverge (separate) after they pass through it.

dragging of inertial frames A colorful name for the effects of gravitomagnetism in which the trajectories of freely-falling bodies are dragged in the same sense as the motion of the source of the gravitational field. *See also* gravitomagnetism, stationary limit.

dust grains Interstellar gas clouds contain not only the basic gas from which stars form, but also dust: solid particles of ice and carbon compounds, which scatter light and obscure distant objects. These particles contribute much of the heavy elements when stars form.

eccentricity A measure of how non-circular an ellipse is. If the ratio of the minor axis (shortest diameter) of the ellipse to its major axis (longest diameter) is r , then the eccentricity is $e = (1 - r^2)^{1/2}$.

eclipse The blocking of the view of one astronomical body by another. Since the Moon and Sun have nearly the same angular size on the sky as seen from the Earth, there are times when the Moon totally blocks the light from the Sun at some locations on the Earth: total eclipses. Similarly, the Earth can come between the Sun and Moon, stopping sunlight from reaching the Moon: a lunar eclipse.

Einstein curvature tensor The mathematical construction used by Einstein to describe the part of the curvature of spacetime that directly equals the densities, momenta, and stresses of the matter producing gravity. The tensor depends on derivatives of the metric of spacetime. *See also* curvature, spacetime, spacetime metric, stress energy tensor, Einstein field equations, derivatives, tensor.

Einstein field equations The fundamental equations of general relativity. On one side of the equation is the Einstein curvature tensor, on the other the stress energy tensor of matter. The equations are differential equations because they contain derivatives of the spacetime metric. There are ten equations in all that must be solved for the ten components of the metric. The equations are non-linear and interlinked, so that in most cases a realistic solution can only be obtained by computer simulation. *See also* Einstein curvature tensor, stress energy tensor, tensor, spacetime metric, component.

Einstein luminosity A number with the dimensions of luminosity (energy per unit time) that is composed only of fundamental constants, c^5/G . It is the maximum luminosity that any physical object can radiate. *See also* luminosity.

Einstein radius The characteristic distance from a gravitational lens at which bright images might appear. If the lens is strictly spherical and the source is directly behind its center, then the image of the source will be a ring at this radius. The ring is called the Einstein ring. *See also* gravitational lensing.

Einstein ring *See* Einstein radius.

electric charge The source of the electric and magnetic field. Particles can have positive, negative, or zero charge. All charges are integer multiples of the fundamental unit of charge, which is the charge on a proton. *See also* magnetic field.

electromagnetism The theory of the electric and magnetic fields, devised by Maxwell in the nineteenth century. A single theory unifying electricity and magnetism is required because moving electric charges create magnetic fields and oscillating magnetic fields create electric fields. *See also* electric charge, magnetic field.

electron neutrinos All leptons, including electrons, fall into three families. Each family contains the lepton, its anti-particle, and an associated neutrino and antineutrino. The numbers of leptons of each type must be conserved in a nuclear reaction. Thus, when an electron disappears, an electron neutrino must be created to take its place. *See also* lepton, anti-matter, neutrino.

electron The fundamental particle that carries the negative electric charge within atoms and which moves through solids to carry electric current. It is a lepton. It is affected by the electromagnetic forces because of its charge and by the weak interaction, but does not sense the strong interaction. *See also* lepton, strong interaction, weak interaction.

electroweak The theory that unifies the electromagnetic and weak interactions. It shows that the weak interaction, which is responsible for beta decay, is related to and of the same strength

as the electromagnetic interaction at very high temperatures. Recent experiments have given very firm confirmation to this theory. *See also* electromagnetism, weak interaction.

electron volt A measure of energy, denoted by eV. One eV equals the energy an electron gains by falling through a voltage difference of one volt. This is $1.6022 \times 10^{-19} \text{ kg m}^2 \text{ s}^{-2}$. *See also* MeV.

elliptical galaxy A galaxy that appears smooth and elliptical in photographs. Some are probably oblate spheroids, but others may be genuinely tri-axial. Astronomers believe that their smooth appearance is caused by the extreme mixing that happens when two less regular galaxies collide and merge.

energy Physicists use the word energy to describe something very specific, not very closely related to the everyday uses of the word. The basic energy is kinetic, equal to $mv^2/2$ for a body of mass m and speed v . All other energies are defined in such a way that the sum of all energies is conserved, unchanging with time. In relativity, the energy includes the rest-mass of the object by the famous formula $E = mc^2$. For a system described by a classical theory of physics, the total energy of a system, including masses, is always positive, but in quantum systems negative energy is allowed, at least for short times. In general relativity, a sufficiently large curvature of time can make the total energy of a particle near a star or black hole negative, but even then the total energy of the particle plus the star must be positive. *See also* conservation of energy, classical, quantum theory, stationary limit.

entropy A measure of the disorder in a system. The larger the entropy, the less structure the system has. It is also a measure of information, since a more chaotic system contains more information: to reconstruct the system exactly would require a larger list of rules than to construct a well-ordered system. The second law of thermodynamics asserts that the total entropy of any closed system cannot decrease with time, and in any realistic system it will increase. Living systems manage to control their own entropy while they are alive, but to do so they must increase the entropy of their environments. Black holes are the objects with the highest known entropy.

epicycles *See* cycles.

ergosphere *See* stationary limit.

Euclidean geometry The geometry that follows the axioms of Euclid, which describe a flat space in which distances follow the usual Pythagorean theorem. A two-dimensional Euclidean geometry has the properties of the surface of a flat piece of paper. We normally assume that we live in a three-dimensional Euclidean geometry, but actually gravity makes tiny changes in the geometry that are not perceptible except over very large regions. *See also* Pythagorean theorem, curvature.

Euclidean plane A two-dimensional Euclidean space. *See also* Euclidean geometry.

events Points of spacetime, having a fixed location and time of occurrence. *See also* spacetime.

excited state In the quantum theory of atomic structure, electrons orbiting the nucleus normally have well-defined values of energy, and the electrons normally occupy all the lowest-energy states, one per state as required by the Pauli exclusion principle. In this configuration the atom is said to be its *ground state*. If an electron has a higher energy than required, so that a lower-energy state is empty, then the atom (as well as the electron) is said to be in an excited state. Left alone, the electron will normally rapidly drop into the lower-energy state, emitting a photon with a characteristic energy and frequency. *See also* quantum theory, atoms, photon, spectral lines.

expansion of the Universe Astronomers have discovered that objects that are more distant from us are systematically moving away from us with a faster speed, proportional to their distance. In such a circumstance, any astronomer anywhere else will also see the same thing, so that the Universe is expanding in a homogeneous way. This expansion can be measured for a variety of different objects independently. If one traces back the expansion in time, then all the observable Universe was compressed into a tiny volume about 14 billion years ago. The expansion at that time was explosive, and has been called the Big Bang. *See also* Big Bang, homogeneous.

experimenter In this book, a complete and careful system for gathering all possible information about events in spacetime. Experimenters carefully synchronize their clocks, they are not fooled into making errors because there is a delay in information reaching the experiment's headquarters from more distant information-gathering stations, they define things like distance and time in exactly the way one would expect, and they can make measurements with arbitrarily good accuracy. Such experimenters are the ones that will observe the unexpected effects of special relativity, such as the Lorentz–Fitzgerald length contraction or the time dilation. They are also called *observers*. *See also* special relativity, Lorentz–Fitzgerald contraction, time dilation.

exponential function The mathematical function describing quantities whose growth rate is proportional to their size. Denoted $\exp(x)$ or e^x , where $e = 2.7182818284$ to ten decimal places. Its inverse is the logarithm function. *See also* logarithmic scale.

fermion Each elementary particle has spin, an intrinsic angular momentum that, because of quantum effects, always is either an integer or half-integer multiple of $h/2\pi$, where h is Planck's constant. Particles that have integer spin are called bosons, those with half-integer are fermions. Fermions are unable to occupy the same quantum state, so they anti-bunch. All the standard elementary particles – electrons, protons, neutrons – are fermions. The fact that electrons cannot occupy the same state is essential for the structure of atoms as we know them, with diffuse clouds of electrons available to other atoms for bonding into chemical compounds. The fact that fermions cannot share a quantum state is called the *Pauli exclusion principle*. *See also* fermion, quantum theory.

finite differences The basis of computer calculations that ap-

proximate the continuous motion of something by a series of small but finite steps. The differences are used in place of the derivatives of calculus. calculus and finite-differences If the steps are small enough, the numerical approximation can be made as accurate as one wishes. *See also* calculus, derivatives.

flat space A space without curvature. This is not quite the same as a Euclidean space. A geometry can be flat without having a metric defined on it, and so without having a Pythagorean theorem. For example, an ordinary chart of, say, temperature in New York against time, is a two-dimensional flat space, but there is no meaning to the “distance” from one point on the curve to another. *See also* Euclidean geometry, curvature, metric tensor, Pythagorean theorem.

flux The term physicists use to denote how much of something is passing through a region in a certain time. The flux of energy is the amount of energy passing through a given surface, per unit area and per unit time. The flux of momentum would be the amount of momentum carried by particles passing through a surface, again per unit area and time. In particular, the energy flux measures what we normally mean by the (apparent) brightness of a source of light: multiplying the flux by the area of the pupil of my eye and by the (small) time it takes the eye to sense light, it tells me how much light I actually see from the source.

frame In relativity, the complete coordinate system that an experimenter (observer) constructs in order to locate events and measure distance and time relations among them. In special relativity, this frame can be a homogeneous coordinate system, with straight coordinate lines and uniform distances and time between them. Such a frame is called an *inertial frame*. In general relativity, the curvature of spacetime prevents large-scale homogeneous coordinate systems from being constructed, so the word *frame* is usually reserved for the local coordinates set up by a locally inertial observer to measure phenomena in a small region around a particular event. *See also* experimenter, homogeneous, inertia.

fundamental frequency The characteristic frequency that has the longest wavelength inside a vibrating body, therefore involving as much of the body as possible in a single coherent motion. *See also* characteristic frequencies, overtones.

galaxy A collection of stars well separated from other such collections and held together by its own gravitational attraction. Most galaxies contain 10^{10} stars or more, as does our own, the Milky Way. But some dwarf galaxies are much smaller, like the Magellanic Clouds. Globular clusters are smaller still, and are not referred to as galaxies, since they are normally part of true galaxies and not isolated on their own. Galaxies usually contain gas as well as stars, and they appear also to contain considerable dark matter, perhaps ten times as much as is luminous. The visible part of the Universe contains some 10^{12} galaxies. *See also* globular clusters.

galaxy cluster Galaxies are not distributed uniformly in space. Instead, they group into loose chains and more tightly bound clusters. The Virgo Cluster is the nearest large cluster, but our

own galaxy belongs to a small cluster called the Local Group. Clusters can have anything from a handful to thousands of galaxies. They also contain dark matter, even more than in the individual galaxies. The statistical distribution of galaxies in clusters and chains provides important information about the way they were formed, and supports the cold-dark-matter model of galaxy formation. *See also* galaxy, dark matter, cold dark matter.

gamma-rays High-energy photons, above about 100 keV, with wavelengths shorter than about 10^{-11} m. This is the highest-energy (shortest-wavelength) section of the electromagnetic spectrum. *See also* infrared, microwave, ultraviolet radiation, X-ray, sub-millimeter.

gamma-ray bursts Every day the Earth receives two or more bursts of gamma-radiation from very distant sources. The bursts are likely produced by either the merger of two neutron stars or a neutron star and a black hole, or by a highly energetic form of supernova explosion called a *hypernova*. If they are mergers, then they will be accompanied by a strong burst of gravitational radiation from the orbital in-spiral that took place before the merger. *See also* gamma-rays, neutron star, black holes, supernova, hypernova, gravitational wave.

geodesic The mathematical name for a curve that follows a locally straight line through a curved space. Always going straight as determined by a locally flat observer, the line can nevertheless wander about because of the curvature of the space. They are therefore good tracers of curvature. On a sphere, geodesics are great circles. In a flat space they are normal straight lines. In spacetime, they are the world lines of freely-falling particles, unaffected by non-gravitational forces. *See also* curvature, flat space, spacetime, world line.

giant In astronomy, a star that has expanded to many times its normal size. This happens when the star exhausts its normal fuel of hydrogen, the central region contracts and heats up, and the star begins to process heavier nuclei into still heavier ones. The core of the star gets denser and its envelope thinner. If the star expands far enough, the outer layers become cool and the spectrum moves into the red. These are *red giants*. Stars that are not quite so big and so are at a higher surface temperature could be *blue giants*. *See also* cataclysmic variable, supergiant, nuclear reactions.

glitches A word that astronomers have adopted to describe rapid changes in the periods of pulsars. They typically show a sudden rapid speed-up followed by a longer slow-down. Their cause is not entirely clear, but it may have to do with the interaction between the superfluid liquid interior and the crust of the star. *See also* pulsar, superfluid, crust.

global Used by mathematicians and physicists to describe concepts that are valid everywhere in a large domain. Its opposite is *local*. As an example, one can say that an observer in special relativity can construct a global frame that is the same everywhere, while one in the curved spacetime of general relativity could construct a similar frame only locally. *See also* local, frame, spacetime, curvature.

globular clusters Roughly spherical, tightly bound star clusters containing hundreds of thousands of members or more. They have a distinctive shape and appearance. Since all their members were formed at the same time, they are good laboratories for learning about relative rates of evolution of different kinds of stars. Encounters among member stars can create binary systems of stars and black holes, and also can cause the cluster to evolve in various ways, and globular clusters may collide with one another or with other concentrations of stars in the Galaxy. They may have been much more plentiful in the past, and indeed perhaps they are among the building blocks of galaxies. *See also* star cluster, galaxy.

grand unified theory Unified theories of the strong interaction with the electroweak. *See also* electroweak, strong interaction, unified field theories.

gravitational collapse The inward fall of a self-gravitating body that can no longer produce enough pressure to resist the pull of gravity. This is the event that triggers supernovae of Type II. *See also* supernova of Type II.

gravitational lensing The action of gravity in bending the path along which light propagates in such a way that images of objects are distorted, duplicated, or changed in brightness.

gravitational slingshot When a small mass, like a spacecraft, encounters a large mass, like a planet, that is moving, the energy of the spacecraft is not the same before and after the encounter. It is possible to arrange the encounter so that the spacecraft gains energy and is slung into a different trajectory. This can enable the craft to reach parts of the Solar System that require more energy than the launch can give it. *See also* conservation of energy.

gravitational wave A ripple in the gravitational field that travels with the speed of light through space. It carries time-dependent tidal accelerations, which are the only ones measurable by a local experiment. The accelerations are transverse to the direction of motion of the wave, and they mimic the accelerations of the masses that produce the waves, as projected on a plane perpendicular to the direction of motion. Because gravity is a very weak force, the waves have only a tiny influence on matter that they pass through. This makes them hard to detect but also makes them good carriers of information, since they do not get distorted by intervening matter. *See also* tidal acceleration, bar detector, interferometer.

gravitoelectric field A term used in this book to describe the dominant part of the gravitational acceleration in general relativity, the part that is embodied in the curvature of time and is generated by the active gravitational mass. For weak fields, this is the ordinary Newtonian gravity. *See also* curvature, active gravitational mass.

gravitomagnetism The part of the gravitational acceleration in general relativity that is generated by momentum, and which acts only on bodies with momentum. It has some resemblance to ordinary magnetism, hence its name. Some physicists call it *magneto-gravity* instead. It is responsible for the dragging of

inertial frames and the existence of stationary limits. *See also* stationary limit, frame, inertia, dragging of inertial frames, momentum.

graviton A quantum of a gravitational wave in the same sense that a photon is a quantum of light. However, while the photon is a well-defined concept grounded in the theory of quantum electrodynamics, the graviton is just a guess: there is no quantum theory of gravity yet. *See also* quantum theory, photon, quantum electrodynamics.

greenhouse effect The trapping of heat by using a material that is transparent to light but not to infrared radiation. Glass in a greenhouse allows sunlight to pass through, where it is absorbed by plants and the ground. The energy is re-radiated as infrared light, because that is where the peak of the black-body spectrum is for typical temperatures on the Earth. The glass is not as transparent in the infrared part of the spectrum as in the visible. Therefore, heat builds up in the greenhouse. This effect is a cause of warming on the Earth. Carbon dioxide, methane, and fluorocarbons are all greenhouse gases, which allow sunlight to reach the ground but which absorb some of the infrared light that is re-emitted back to space. This effect has existed for billions of years; without it the Earth would be much colder. The concern today is that human activity is raising the amount of greenhouse gas and hence the mean Earth temperature beyond safe levels. *See also* infrared, greenhouse gases.

greenhouse gases Gases that create a greenhouse effect by being transparent to visible light but at least partly absorbing at the longer infrared wavelengths. *See also* greenhouse effect.

half-life The time it takes for half of the members of a sample of radioactive particles to decay. Each particle decays at random and has no “memory” of how long it has been waiting to decay. Given a radioactive particle that was created a million years ago, the probability that it will decay in the next year is the same as the probability that it would have decayed in its first year. It follows that, if half the particles decay in a certain time, then half of the remaining ones will decay in the same period of time again. This is the half-life of the particles.

Hawking radiation The black-body radiation emitted by a black hole. This emission is a purely quantum effect: in classical general relativity, black holes cannot emit any radiation. *See also* black-body radiation, black holes, quantum theory, classical, Hawking temperature.

Hawking temperature The temperature of the black-body radiation emitted by a black hole. If the hole is spherical, this is inversely proportional to the mass M of the hole. *See also* Hawking radiation.

helioseismology The study of the characteristic frequencies of oscillation of the Sun. Thousands of frequencies have been measured, and this tightly constrains the solar model. In this way, astronomers “see” deep inside the Sun. This information has helped to point the way to solutions of the solar neutrino puzzle. *See also* characteristic frequencies.

Hertzsprung–Russell diagram A chart plotting the luminosity and temperature of a number of stars. Stars do not appear at random locations in this plot. Most fall in a narrow band called the *main sequence*. Others fall in regions called the giant branches. White dwarfs are located in another small region. These groups show that, for a star with a given mass and at a given stage of its evolution, the temperature and luminosity are related to one another. *See also* main sequence stars, giant, white dwarf.

homogeneity/isotropy problem The puzzle of why the Universe shows such uniformity in all directions and at all distances from us. In the standard Big Bang model, regions sufficiently far from us in different directions have not had time to make contact before emitting the light we see from them. They could not, therefore, have come to some kind of equilibrium, and yet they look very similar. *Inflation* solves this problem. *See also* homogeneous, isotropic, Big Bang, inflation.

homogeneous The same everywhere. The Universe is homogeneous at a given moment of cosmological time, provided one averages over volumes containing many clusters of galaxies. *See also* galaxy cluster.

horizon The outer boundary of a black hole or the limit of what we can see in cosmology. For black holes, the surface is technically known as the *event horizon*, and it is defined as the boundary between events that can send light rays to a very distant observer and those that cannot. For cosmology, the surface is technically known as the *particle horizon*, and it is defined as the boundary between events that could send light to us and those that could not, since the Big Bang. (We ignore scattering and absorption here, just asking whether a particle traveling at the speed of light could reach us.) The event horizon expands if matter falls into the black hole; the particle horizon expands all the time, since each moment we can see some regions that had, until then, been too far away. *See also* black holes, Big Bang.

hot dark matter Dark matter that is composed of particles whose masses are so small that, at the time and temperature of decoupling, the velocities of the particles were too large to allow fluctuations in their densities to grow fast enough to form the seeds of galaxy formation. This mass is about 10 eV. Hot dark matter particles would today be distributed much more smoothly in the Universe than is the visible matter. *See also* decoupling, dark matter, cold dark matter.

Hubble constant The present relative rate of expansion of the Universe, that is the ratio of the speed of recession of a distant galaxy to the distance of the galaxy. This ratio is the same for all galaxies, no matter where the observer stands, in a perfectly homogeneous universe: hence the word “constant” in the name. The real Universe is not perfectly homogeneous, and so measuring the Hubble constant has not been easy. Only very distant galaxies give good values, since their random motions are small compared to the systematic expansion speed. But it is difficult to estimate the distances to such galaxies. Astronomers seek *standard candles*, objects whose luminosity is known so that their distance can be estimated from their apparent bright-

ness. Recently Type Ia supernovae have become useful standard candles. *See also* cosmological scale-factor, homogeneous, standard candle, supernova of Type Ia.

Hubble time The time it would have taken the Universe to reach its present size if its expansion were constant in time. This is the reciprocal of the Hubble constant. *See also* Hubble constant.

hyperboloid A three-dimensional geometric figure obtained by rotating a hyperbola about its axis of symmetry.

hypernova An unusually powerful supernova resulting from gravitational collapse, as does a supernova of Type II. *See also* gravitational collapse, supernova of Type II.

hypotenuse The long side of a right-angled triangle, opposite the right angle.

indices Labels attached to components of vectors or tensors to indicate which directions the components are associated with. *See also* component, Pythagorean theorem.

inertia Essentially, the mass of a body. The word is often used to describe the property of a body that makes it resist acceleration and attempt to continue moving in a straight line. The concept is a little vague, and is made much more precise by Newton's laws of motion. Inertia is simply a property: something either has it or it does not. By contrast, mass is a quantity: it is measured in kilograms in the SI system, and it is meaningful to say that one body has twice as much mass as another. In special relativity, the word is applied to the special coordinate system, or frame, that should be used by experimenters: an *inertial frame*. *See also* mass, frame.

inflation In cosmology, the postulated period of time during which the very early Universe expanded exponentially rapidly. Such a phase, if it occurred, would explain the homogeneity of the Universe and many other observed properties. *See also* negative pressure.

infrared Region of the electromagnetic spectrum extending from the red end of the visible spectrum to longer wavelengths, typically from $0.7\text{ }\mu\text{m}$ to about 1 mm. *See also* microwave, X-ray, gamma-rays, ultraviolet radiation, sub-millimeter.

inhomogeneity Non-uniformity, condition of not being homogeneous. *See also* homogeneous.

innermost stable circular orbit A unique feature of orbits around black holes and other ultra-compact objects, which is not present in Newtonian gravity, is that there is an inner limit to circular orbits; inside this limit, circular orbits exist but are unstable: any small disturbance or non-circularity will make the orbit diverge rapidly from the original circular form. These orbits set limits on how far inwards an accretion disk can extend. *See also* accretion disk.

interferometer An instrument designed to measure with great sensitivity changes in the difference between two lengths. The lengths are called the arms of the instrument, and the measurement technique is to split coherent light along the two arms, reflect it from the ends and look at the interference pattern

formed when the light re-combines. Changes in the difference between the arm-lengths change the pattern. Interferometers are used as gravitational wave detectors by making the ends of the two perpendicular arms free: when a gravitational wave comes along, it changes the lengths of the two arms in different ways, thereby creating a signal at the output. *See also* bar detector.

interstellar clouds Dense clumps of gas and dust in the Galaxy, where star formation occurs.

invariance Independence, the condition of being unchanged when something else changes. In physics, this is used to describe systems that have a symmetry. A system that is independent of time is time-invariant. If it is symmetrical under rotations, it is rotation-invariant. In dynamics, systems that are invariant have associated conserved quantities. The time-invariance of the spacetime of special relativity (Minkowski spacetime) insures that physical systems in special relativity have a conserved energy. The spherical symmetry of the non-rotating black hole (Schwarzschild solution) insures that particles orbiting the hole have a conserved angular momentum. *See also* conservation of angular momentum, conservation of energy, Minkowski spacetime, black holes.

invariant hyperbola The set of all events in the spacetime of special relativity (Minkowski spacetime) that have a fixed interval from the origin. This definition is independent of the observer, so the set is invariant under a change of observer. The set forms a hyperbola when drawn in just two dimensions, t and x , say, where the equation is $c^2 t^2 - x^2 = k$ for some fixed k . *See also* Minkowski spacetime, spacetime-interval.

inverse-square Depending inversely on the square of a variable. In gravitation, the Newtonian gravitational force is proportional to $1/r^2$, so it is an inverse-square law in the distance r from the source of gravity.

ionized Having lost one or more electrons (said of atoms). An atom that has all of its electrons is charge-neutral. If one or more are removed, it has a positive charge and is called a positive ion.

Irregular galaxy A galaxy that is not classifiable as either spiral or elliptical. *See also* spiral galaxy, elliptical galaxy.

isothermal Literally, of uniform temperature. A gas that keeps its temperature constant when it expands or compresses is said to have an isothermal equation of state.

isotope Atoms of a given element must all have the same number of protons in their nuclei, and (unless ionized) the same number of electrons in orbit around the nucleus; but they do not have to have the same number of neutrons in the nucleus. The chemistry of the element depends on its charged particles; neutrons are important only because they help to hold the nucleus together by the strong interaction. Atoms of the same element that have different numbers of neutrons are said to be different *isotopes* of the element. *See also* ionized, chemical elements, nucleus, proton, electron, deuterium, strong interaction.

isotropic The same in all directions from a given point. A sphere is isotropic about its center, but not about other points. The Euclidean plane is isotropic about all points. If something is isotropic about all points it must also be homogeneous, and if something is homogeneous and also isotropic about one point, it must be isotropic about all points. *See also* anisotropy, homogeneous, Euclidean plane.

Jeans length The scale on which gravity overwhelms pressure in a homogenous gas. Small disturbances in a homogeneous gas on very short length-scales will bounce back and smooth out because of gas pressure, but on longer scales the attraction of gravity causes them to grow. This is the Jeans instability. It plays a role in star formation, and it may have been important in galaxy formation as well. The amount of gas enclosed in a sphere whose radius was the Jeans length at the time of decoupling was approximately the mass of a typical globular cluster today; this suggests that globular clusters formed as gas instabilities in the expanding universe. *See also* homogeneous, globular clusters.

jets Narrow linear plume of gas moving away from an astronomical object at high speed. A great many objects are seen to produce jets on different length-scales: newly forming stars, pulsars, black hole systems, radio galaxies, quasars, and more. *See also* pulsar, quasars.

joule SI unit of energy, equal to the work done by a force of one newton moving through one meter, or twice the kinetic energy of a one-kilogram mass moving at 1 m s^{-1} .

kelvin The SI scale for absolute temperature. It has the same degree size as the standard Celsius scale, but its zero is at absolute zero. *See also* Celsius.

kiloparsec One thousand parsecs, or about $3.085678 \times 10^{19} \text{ m}$. *See also* megaparsec, parsec.

Kuiper Belt The region outside the orbit of Neptune which seems to contain a large number of planetesimals, which is a reservoir from which large asteroids occasionally fall toward the Sun. It is thought to be a relatively thin ring, having been formed during planetary formation. *See also* planetesimal, Oort Cloud.

laser A source of coherent light, which is light whose photons have the same frequency and phase. Its creation depends on the fact that photons are bosons. A laser needs a “pump”, which is the source of energy for the light and which arranges a large number of atoms in the laser to be in an excited state. When one of these atoms spontaneously decays to its ground state, emitting a photon, this photon will induce other atoms to decay and emit photons of exactly the same frequency and phase. This induced emission is a purely quantum effect. *See also* photon, boson, excited state, quantum theory.

latitude The usual measure of North–South position on the Earth, running from 90° South (or -90°) at the South Pole to 90° North (or 90°) at the North Pole. *See also* co-latitude, longitude.

law of sines The geometrical relation in a triangle in which the ratio of the sine of any angle to the length of the side opposite that angle is the same for all three angles.

laws of motion The three statements that Newton formulated as a sufficient set to determine the motions of bodies when forces are applied to them. The first law says that, if there is no net force on a body, it will move in a straight line. The second law says that the acceleration of a body is proportional to the net applied force divided by the body’s mass. The third says that if one body exerts a force on another, then the second exerts an equal and opposite force on the first. These laws formed the basis of the study of mechanics until Einstein. *See also* mechanics.

lepton Literally, “light particle”: particles that are affected by the weak interactions and (if charged) the electromagnetic force, but not by the strong interactions. Three kinds of leptons are known: electron-leptons, mu-leptons, and tau-leptons. Each is named for its “meson”, and each family consists of the meson, an associated neutrino, and the anti-particles of these two. *See also* baryon, neutrino, anti-matter, weak interaction, strong interaction, lepton number, muons.

lepton number The net number of leptons in a system or reaction, with anti-particles counting negatively. There are actually three kinds of lepton numbers, for the three kinds of leptons: electron leptons, mu-leptons, and tau-leptons. *See also* lepton, weak interaction, muons.

light-cone In spacetime, the set of events that can be connected to a given event by a single null line, a line along which light could travel. *See also* spacetime.

lightlike In spacetime, a separation between two events is *lightlike* if the events can be connected by a line along which light can travel. *See also* light-cone, spacelike, timelike.

linear Described by a straight line. In mathematics, the relation between two variables y and x is *linear* if the equation relating them has the form $y = mx + b$, for constant m and b . *See also* non-linear.

local Used by mathematicians and physicists to describe concepts that are valid only sufficiently near a particular point. Its opposite is *global*. Smooth geometries are said to be *locally flat*. *See also* global, locally flat.

locally flat The property that smooth geometries have, that they can be approximated very well by a flat geometry in a sufficiently small region around any point, as is familiar by the fact that street-maps printed on flat sheets of paper work well within cities. They are said to be locally flat at that point. *See also* local.

logarithmic scale The scale on a graph in which the markings are separated by an amount proportional to the logarithm of the quantity being displayed. This implies that marks at uniform steps represent uniform factors in the increase of the number, not uniform steps in size. Typically these may be shown as factors of ten steps for each mark. Such a scale is useful for showing the structure of curves that change by large amounts over

their range, and for showing exponential and power-law relationships between variables, since these plot as straight lines on graphs where one or both axes have logarithmic scales, respectively. *See also* power-law, exponential function.

longitude The usual measure of East–West position on the Earth, running from 180° West (or -180°) in the Pacific Ocean through 0° at Greenwich, England, to 180° East (or 180°) in the Pacific again. *See also* latitude.

longitudinal In the theory of waves, a *longitudinal* wave is one whose action is along the direction of its motion. Sound waves are longitudinal, whereas water waves are transverse. *See also* transverse.

loop In computer programs, a group of instructions that is executed repeatedly.

Lorentz–Fitzgerald contraction The change in length of a moving body in special relativity.

Lorentz–Fitzgerald transformation The rule for changing from the spacetime coordinate system (t, x, y, z) of one experimenter in special relativity to that of another. *See also* experimenter.

loss of simultaneity The fact that, in special and general relativity, there is no universally agreed notion of whether or not two different events occurred at the same time. One experimenter, measuring as carefully as possible, might assign the same time-of-occurrence to two events. A different experimenter, one who is moving along the direction separating the two events, and who is making the same set of measurements and using the same definition of simultaneity, will place the event that is toward the front in the direction of his motion at an earlier time than the other one.

luminosity The amount of energy radiated by an object per unit time. This is the intrinsic brightness. The apparent brightness depends on how far away the object is. Astronomers measure luminosity in absolute magnitudes. *See also* absolute magnitude, apparent magnitude, Einstein luminosity.

macroscopic The word physicists use to denote aspects of the world that are on a large enough scale to be perceived by the eye or other senses; opposite of microscopic.

magnetars Neutron stars with ultra-strong magnetic fields, seen as pulsars with very long pulse periods. *See also* magnetic field, neutron star, pulsar.

magnetic field Magnetism is the force created by moving charges that is not present if the charges are at rest. Magnetism acts only on moving charged particles, and the force is proportional to their charge and their speed of motion. The *magnetic field* is the term used for the magnetic force on a particle per unit charge and per unit speed, i.e. the part of the magnetic force that depends just on the particles that create it. The field extends through all space, but it only exerts a force wherever a moving charged particle may be. (This is what physicists mean by the word *field*.) The field has an energy spread out through space, and it can change with time, carrying waves: electromagnetic waves. Similar remarks apply to the electric field, which is

created by charges and acts on charges, regardless of their state of motion. *See also* electric charge, electromagnetism.

magnetic monopoles Hypothetical particles with a magnetic charge, that is carrying just a North magnetic pole or a South magnetic pole. No such particles have been discovered, but there is reason in theories of high-energy physics to believe that they may have been abundant in the very early universe. They would behave like electric charges but with magnetic and electric fields interchanged: a static magnetic monopole would create a magnetic field, while a moving magnetic monopole would create an electric field. The theory of inflation explains why they are so rare now that they have not been seen. *See also* electric charge, magnetic field, inflation.

magnitude Normally used as shorthand for *apparent magnitude*. *See also* apparent magnitude.

main sequence stars Stars that are burning hydrogen in their cores to power their luminosity. The term refers to the narrow sequence of points in the Hertzsprung–Russell diagram occupied by such stars, forming a nearly one-dimensional sequence according to the mass of the star. When stars begin to burn heavier elements they become giants, and eventually evolve either to white dwarfs or to supernovae. *See also* Hertzsprung–Russell diagram, giant, white dwarf, supernova.

maser The radio-wave analog of a laser. Masers were invented in laboratories before lasers, and even earlier by Nature: many dense molecular clouds, stars, and accretion disks radiate masers. The radiation comes from transitions in molecules, which typically have much less energy and therefore longer wavelength than internal transitions in atoms. Masers have very small size and, since the wavelength of the radiation is known from measurements on molecular transitions in the laboratory, they allow astronomers to follow the motions of very precisely located regions of gas clouds. They have helped to measure the masses of some supermassive black holes. *See also* black holes, laser, molecules, accretion disk, excited state.

mass The substance of a body; its resistance to acceleration: the more mass an object has, the smaller will be its acceleration in response to an applied force. In relativity, energy has mass ($m = E/c^2$) and resists being accelerated. *See also* weight, inertia.

mass function A particular function of the masses of two stars in a circular binary orbit, and of the angle of inclination of their orbit, which is measurable if it is possible to follow the speed of one of the stars along the line-of-sight to the system. This information is all that is normally available for most binary systems observed optically. *See also* angle of inclination, spectroscopic binary.

mass-to-light ratio The ratio of the mass of an astronomical system to its luminosity, in units of the solar values: $(M/M_\odot)/(L/L_\odot)$. From studying many kinds of stars and galaxies, astronomers have a rough idea of what this ratio should be for a typical system. This allows them to estimate the mass of the system from a measurement of its luminosity. *See also* luminosity.

matrix A mathematical structure having the form of an array comprising rows and columns.

matter-dominated In cosmology, a cosmological model in which the dynamics is governed by matter rather than radiation or dark energy. In practice this means that pressure is negligible in determining the evolution of the cosmology. *See also* cosmology, dark energy.

mean Mathematical term for the average of a set of numbers.

mechanics The study of the motions of bodies in response to applied forces.

megaparsec One million parsecs, or about 3.085678×10^{22} m. *See also* kiloparsec, parsec.

metric tensor The mathematical structure that allows one to calculate distances in a curved space, or intervals in a curved spacetime. In any particular coordinate system it can be represented by a matrix of values, which may change from place to place. *See also* matrix, spacetime-interval.

MeV One million electron volts of energy, or 1.6022×10^{-13} kg m² s⁻². It is frequently used as a measure of mass; the energy equivalent of the mass of an electron is about 0.5 MeV. *See also* electron volt.

microlensing Gravitational lensing phenomenon in which the lens is a star rather than a galaxy. Since individual stars are very much smaller than galaxies and their random velocities are much higher, microlensing tends to be a short-lived phenomenon. It can give very useful information about the sizes of light-emitting regions in the object being lensed. *See also* gravitational lensing.

micron Another name for a micrometer, 10^{-6} of a meter.

microwave An electromagnetic wave with a wavelength longer than infrared; the beginning of the radio region of the spectrum. Typical wavelength range is 1 mm to 1 m. *See also* infrared, ultraviolet radiation, X-ray, gamma-rays, submillimeter.

millisecond pulsars Pulsars with periods shorter than 10 ms. *See also* pulsar.

Minkowski spacetime Spacetime with no gravitational effects; flat spacetime. *See also* spacetime, curvature.

missing mass Matter that is inferred to be present in galaxies, clusters, and the Universe as a whole by the fact that the dynamics of these systems cannot be explained by the masses of observed stars. The missing mass is some 100 times as much as the luminous mass. Some of it may be in ordinary gas, but most of it must be in some form of matter that has not yet been observed experimentally. *See also* cold dark matter, hot dark matter, galaxy cluster.

molecules Systems of atoms joined together by the mutual attractions of the electrons of the atoms for the nuclei of other atoms in the molecule. Molecules are the building blocks of chemicals, and the study of the combinations of atoms into molecules is the main subject of the science of chemistry. In

astronomy molecules are formed in the cool outer regions of giant stars and in molecular clouds. *See also* atoms, nucleus, electron, giant.

momentum The product of the velocity of an object with its mass. Like angular momentum, this is conserved in Newtonian mechanics and special relativity, and in general relativity if the spacetime is invariant under translations in the direction of the momentum. *See also* invariance, spacetime.

muons One of the types of lepton mesons. Along with electrons and tau-mesons, muons form one of the three families of leptons. It has associated with it the mu-neutrino. *See also* lepton, lepton number, electron.

naked singularities Singularities in general relativity that are not hidden within a horizon but are visible to other parts of the universe. A naked singularity would represent the breakdown of the predictive power of general relativity, and would presumably mean that the theory had to be replaced. Serious singularities of this type are not known, and are postulated not to exist (cosmic censorship hypothesis). This has not yet been proved. *See also* singularity, cosmic censorship hypothesis.

nebula A diffuse cloud of gas around a star. Originally astronomers used this term for any diffuse clouds of light on the sky. As telescopes improved, some nebulae turned out to be star clusters in our Galaxy, others external galaxies. This obsolete usage is preserved in some traditional names, like the "Great Nebula in Andromeda" for the Andromeda Galaxy M31. *See also* star cluster, galaxy, planetary nebula.

negative pressure Normal gas pressure is positive, in that it pushes outwards on the walls of its container. Systems have negative pressure if they pull in on their containers. A stretched rubber band has negative pressure along its length; this is called *tension*. Since pressure contributes to the gravitational field through the active gravitational mass, a sufficiently large negative pressure can turn the gravitational field repulsive. This is the explanation for inflation. *See also* active gravitational mass, inflation.

neutrino Leptons of very small or zero mass that are produced in beta decay and many other nuclear reactions. There are three kinds of neutrinos, associated with each of the three kinds of leptonic mesons. The flux of neutrinos from the Sun is smaller than expected, and this has led to revisions in the theory of leptons and the determination that neutrinos of one kind seem to transform themselves into one another. *See also* lepton.

neutron The electrically neutral particle that, with the proton, is one of the building blocks of the atomic nucleus. Its mass is slightly larger than that of the proton, large enough that a single isolated neutron can decay into a proton, an electron, and a neutrino. This process of beta decay can also happen inside a nucleus that has too many neutrons. Neutrons are stable in nuclei that don't contain too many of them, and they are stable within neutron stars, of which they are the main constituent. *See also* baryon, beta decay, neutron star, proton, strong interaction.

neutron star A star whose support against gravity comes from the pressure of a degenerate gas of neutrons. Its typical mass is about a solar mass, and its radius about 10 km. Some neutron stars are pulsars, but there are probably many more that are not detected through any emitted radiation. *See also* degenerate gas, neutron, pulsar, white dwarf.

non-linear A mathematical term for a relationship between two variables that is not linear, so it does not plot as a straight line. *See also* linear.

normal modes The patterns of vibration of an object that are associated with its characteristic frequencies. For each frequency there is a specific kind of motion at different points in the body, and this pattern is the normal mode associated with that frequency. *See also* characteristic frequencies.

nova A star that suddenly brightens up. The change in brightness is not as great as for a supernova. Where a supernova represents the destruction of a star, a nova is caused by changes in the rate of accretion onto a compact star, and therefore can recur in the same system. *See also* accretion, supernova.

nuclear reactions Reactions involving changes in the composition of nuclei, such as combining two protons and two neutrons to make a helium nucleus, or combining three helium nuclei to form a carbon nucleus. While chemical reactions involve only the electrons and leave the nuclei intact, nuclear reactions change the nuclei and therefore the element. They require higher temperatures, energies, and densities than chemical reactions. Reactions that release energy constitute the main source of the energy radiated by stars.

nucleon A neutron or proton: the constituents of the nucleus. Nucleons exert forces on one another via the strong interaction, which electrons and neutrinos do not feel. *See also* neutron, proton, strong interaction.

nucleus The positively charged center of an atom, containing protons and neutrons. The number of protons determines the element that the atom belongs to, and its chemical properties. *See also* chemical elements, neutron, proton.

observer *See* experimenter.

Occam's razor The principle that any hypothesis or theory devised to explain a new phenomenon should be as simple as possible and involve as few new assumptions and undetermined parameters as possible.

Olbers' Paradox The question: why is the sky dark at night? In an infinitely large and infinitely old universe filled uniformly with stars, the sky would be infinitely bright, and our Sun would not even be noticeable. Our Universe must be either of finite age or of finite size, or both.

Oort Cloud The roughly spherical cloud in which comets originate, far outside the orbit of Pluto and outside the Kuiper Belt. This is thought to have been left over from the formation of the Solar System; at such large distances, where light from the Sun is very weak, icy comets formed but never evolved into the planetesimals that inhabit the closer Kuiper Belt. *See also* Kuiper Belt, planetesimal.

overtones Characteristic frequencies that are higher than the fundamental frequency. For a simple stretched string, the overtones are at integer multiples of the fundamental, but in more complicated systems they are not. *See also* characteristic frequencies, fundamental frequency.

ozone The compound O₃, made of three oxygen atoms. It is only very weakly bound, and can be split up by the addition of a small amount of energy. It forms in the upper atmosphere of the Earth where collisions among molecules are infrequent enough that the molecule can have a long lifetime. It is a good absorber of ultraviolet light from the Sun, and protects living things from this damaging radiation. Man-made pollutants, particularly fluorocarbons, have reduced the concentration of ozone dramatically at some latitudes.

panspermia The idea that life could have originated somewhere else in the Universe and come to the Earth soon after it formed. The evolutionary record would not have been different, but the initial primitive living organisms would have come from somewhere else.

parallax The apparent change of position on the sky of an astronomical object, caused by the Earth's orbital motion around the Sun. Objects near the Earth seem to move back and forth relative to objects more distant. The amount of apparent motion of a near object measures its distance.

parsec The standard unit of distance in astronomy, about 3.085678×10^{16} m. It is defined as the distance to a star whose parallax is exactly one second of arc. It equals 3.26 light-years. *See also* megaparsec, kiloparsec.

particle horizon *See* horizon.

pattern matching The process of searching through a set of data to find something that matches a pre-determined pattern. The detection of gravitational waves relies on pattern matching, because the strength of the waves is not great enough to make them visible in the raw data output from a detector. Instead, scientists look for disturbances that, systematically over time, match a predicted waveform.

Penrose process A method of extracting energy from a black hole by making use of negative-energy orbits inside the stationary limit. If a positive-energy particle falls into the stationary limit and splits into two, one of which has negative energy, then when the other one escapes from the hole it will carry more energy than it began with. This energy comes from the rotation of the hole. *See also* stationary limit.

periastron *See* perihelion.

perihelion The point on an elliptical orbit around the Sun that is nearest to the Sun. If the central object is a star, the point is called the *periastron*; if the Earth, *perigee*. *See also* aphelion.

perpetual motion The idea that an isolated, realistic physical system could somehow execute a particular motion indefinitely. Since all real physical processes involve some kind of friction or dissipation, perpetual motion requires that the energy be replenished, and this conflicts with the principle of conservation

of energy for an isolated system. No such systems have ever been found. *See also* conservation of energy.

photoelectric effect The ejection of an electron from a metal when light of a certain frequency falls on it. If the light has too low a frequency, nothing happens. The higher the frequency of light beyond the critical value at which electrons begin to be ejected, the greater the energy of the ejected electron. This was explained by Einstein to be a consequence of the relation between energy and frequency for photons. Once a photon has enough energy to tear the electron away from the metal, any extra energy (on account of the larger frequency) goes into the kinetic energy of the electron. *See also* photon.

photon A concept introduced by Einstein and now fully explained by the theory of quantum electrodynamics. According to this picture, light sometimes behaves like a particle whose energy E is proportional to its frequency f , $E = hf$, where h is Planck's constant. *See also* quantum electrodynamics.

photosphere The outer layer of the Sun from which comes the light that we see. Light scatters a huge number of times on making its way outwards from the Sun's central region, so the photosphere is the surface of last scattering.

physical cosmology The study of the physical processes in the early Universe, including the formation of elements, the first stars, and galaxies.

pixel The smallest area in an image that can be resolved, i.e. distinguished from neighboring areas. In a photograph, this would be the size of a grain of the emulsion, the smallest unit that is exposed by light. In a digital camera or video camera, this is the size of one element of the CCD that is used as the sensing device.

Planck length A number with the dimensions of length which is formed purely from the fundamental constants. Its value is $(\hbar G/c^3)^{1/2} = 4 \times 10^{-35}$ m. *See also* Planck mass, Planck time.

Planck mass A number with the dimensions of mass which is formed purely from the fundamental constants. Its value is $(\hbar c/G)^{1/2} = 5.5 \times 10^{-8}$ kg. *See also* Planck length, Planck time.

Planck time A number with the dimensions of time which is formed purely from the fundamental constants. Its value is $(\hbar G/c^5)^{1/2} = 1.4 \times 10^{-43}$ s. *See also* Planck length, Planck mass.

plane of the sky The "sky" is the astronomers' word for the celestial sphere, which is a two-dimensional sphere around the Earth; all distant objects are projected onto this sphere to get their angular location on the "sky". Near any particular point, the celestial sphere may be approximated by a plane, because it is locally flat. This is the plane of the sky at that point. *See also* locally flat.

plane wave A wave that propagates with a planar wave-front. All waves that spread out from a localized source are effectively plane waves far away, when the distance over which the curvature of their wave-front is noticeable is much larger than an experimenter's apparatus.

planetary nebula The shell of gas expelled by a giant star during its transition to becoming a white dwarf. These are among the most beautiful objects in the Galaxy when photographed with sufficient resolution.

planetesimal A rocky fragment of the kind that accumulated into planets. Individual examples remaining today are asteroids. *See also* asteroids, Kuiper Belt.

plasma A gas consisting of free electrons and ions. It must be hot enough to prevent the ions and electrons from recombining into neutral atoms.

plate tectonics The process by which the continents have moved around the Earth.

point mass The idealization of a simple elementary particle as a point. This model cannot be realistic, since if the particle is charged the electric field would be infinitely large, and the energy required to assemble the particle would also be infinite. String theory attempts to remedy these problems by representing particles as two-dimensional loops. *See also* string theory.

polarization The direction, or set of directions, in which a wave acts. A longitudinal wave has only one polarization, along the direction of motion. But a transverse wave can be polarized in various ways in the transverse plane. An electromagnetic wave acts along a line, so there are two independent polarizations along the two perpendicular axes. A gravitational wave acts with ellipses, and these have two independent orientations in the transverse plane, rotated by 45° from one another. *See also* longitudinal, transverse.

polytrope In astronomy, a stellar model constructed using a power-law relation between pressure and density. *See also* power-law.

position vector The vector that locates the position of an object; it stretches from the origin to the location of the object. *See also* vector.

positron The anti-particle of the electron, with the same mass but a positive electric charge. *See also* anti-matter, electron.

post-Newtonian The name used for an approximation to general relativity which describes systems that have weak gravitational fields as basically Newtonian systems with corrections. The corrections are called post-Newtonian terms.

power The rate of doing work, or the rate of expending energy.

power-law A mathematical term for a relationship between two variables in which one is proportional to the other raised to a constant power.

primordial black holes Black holes formed in the very early Universe. Normal stellar evolution leads only to black holes larger than about a solar mass. The conditions to form smaller black holes do not exist today, because the required density is much larger than nuclear. However, in the early Universe, when the average density was larger than nuclear, density irregularities could conceivably have collapsed to black holes of smaller mass. These could contribute to the dark matter today, except that any holes smaller than about 10^{12} kg would have

decayed by now due to the Hawking radiation. *See also* black holes, gravitational collapse.

principle of general covariance One of the ideas that guided Einstein's development of general relativity. It is the statement that the equations describing gravity and matter fields must take the same form in any coordinate system. No system is to be preferred, unlike the situation in special relativity, where inertial frames are singled out as the ones that should be used by experimenters. *See also* frame, inertia.

principle of mediocrity *See* Copernican principle.

principle of relativity The principle that the laws of physics should be independent of the motion of the experimenter who tests them. This was first enunciated in respect to gravity on the Earth by Galileo. Einstein made it a cornerstone of his special relativity. He later generalized it to the principle of general covariance when he created general relativity. *See also* principle of general covariance, experimenter.

proper distance The distance as measured by a local experimenter; independent of coordinate system.

proper time The time as measured by a local experimenter's clock; independent of coordinate system.

proton The fundamental positively charged particle, which is one of the building blocks of the nucleus of all atoms. *See also* atoms, baryon, electron, neutron, nucleus.

protostar The collapsing cloud of gas, on its way to forming a star, which is radiating light because of the release of gravitational energy, but within which nuclear reactions have not yet begun. *See also* nuclear reactions.

pulsar A spinning neutron star which emits a beam of radiation that sweeps the sky as the star turns. When observed from Earth, if the observer is in the beam, the radiation pulses on and off. Most pulsars emit radio waves, some are observed to pulse in optical light or even X-rays and gamma-rays. The beams appear to be formed at the poles of strong magnetic fields, but the mechanism is not understood. Most known pulsars spin several times per second; some spin several hundred times per second. *See also* neutron star, X-ray, gamma-ray, magnetars, millisecond pulsars.

Pythagorean theorem The theorem that gives the length of the hypotenuse c of a right triangle in terms of the other two sides a and b : $c^2 = a^2 + b^2$. The relation defines the metric of Euclidean space. *See also* hypotenuse, metric tensor.

quadratic equation In mathematics, an equation containing the square of an unknown variable, but no higher powers.

quadrupole formula The expression giving a first approximation to the gravitational radiation emitted by a system with weak internal gravitational fields.

quanta Discrete amounts of something.

quantum electrodynamics The quantum theory of the electromagnetic field. *See also* quantum theory, electromagnetism.

quantum fluctuations In quantum electrodynamics, the electromagnetic field undergoes fluctuations that would not be allowed in classical theory, where fields have perfectly well-defined values. The fluctuations lead to a number of effects, such as the Hawking radiation from black holes and a possible explanation for the cosmological constant. *See also* quantum electrodynamics, cosmological constant, Hawking radiation.

quantum gravity The hoped-for theory that will generalize general relativity to a quantum theory of the gravitational field. Most physicists expect that this will happen only through unifying gravity with other forces, as in string theory. Others hope to find a theory of the quantum gravitational field alone. *See also* quantum theory, string theory.

quantum theory A theory of physics that incorporates the characteristic features of quantum phenomena: uncertainty in measurements, radiation fields behaving sometimes like waves and other times like particles, predictions only of the probabilities of the outcomes of experiments rather than certainties.

quark matter Matter so dense that nucleons overlap and their constituents, the quarks, are the true particles of the gas. *See also* quarks.

quark soup *See* quark matter.

quarks Baryons are not the most elementary of particles. They are composed of three building blocks, called quarks. Quarks have a remarkable interaction among one another: it is impossible to pull a quark away from others and isolate it. *See also* baryon.

quasars The brightest continuous light sources in the Universe. They seem to be driven by accretion onto a supermassive black hole. Most are at great distances, which suggests that the ones in our neighborhood have died away. *See also* accretion.

quintessence A word used by some physicists for theories of dark energy that explain why we are seeing an acceleration in the expansion of the Universe today. *See also* dark energy.

radians The mathematical measure of angles that is more natural than degrees. The size of an angle in radians is the ratio of the length of an arc to its radius. This measure runs from 0 to 2π for a full circle.

radiation reaction The force on a system that is created by the radiation that the system emits. Any loss of energy must be reflected in a force that opposes the motion. This is a self-force, created by the particle's own field.

ram pressure The pressure exerted by a stream of gas when it encounters a wall; this depends on its speed and density.

redshift The lengthening of the wavelength of a wave. This can happen because of the motion of its source or its receiver, or because of gravity.

relaxation time The time it takes for a system to reach a form of equilibrium. For clusters of stars, this is the time to share out the energy of a perturbation among the stars, losing any memory of the original perturbation.

relaxed A cluster is relaxed if the distribution of its stars and their velocities bears no memory of the history of the cluster and the origin of its stars.

rest-mass The mass of a body when it is at rest, as inferred from Newton's second law. *See also* laws of motion, mass.

rotation curve The graph of the rotational speed of the stars and gas in a spiral galaxy against their distance from the center. From this it can be inferred how much mass the galaxy has and how it is distributed with radius. These curves have revealed that there is a great deal of mass outside the region of such galaxies which emits light. *See also* dark matter.

scalars The mathematical term for an ordinary quantity that is not associated with any direction. Temperature, density, and pressure are scalars. The position and velocity of an object are, on the other hand, vectors. *See also* vector.

scale-height A term used for a length that is typical for the change in some quantity. *See also* Jeans length.

Schwarzschild radius The radius of the horizon of a Schwarzschild black hole, equal to $2GM/c^2$, where M is the mass of the black hole. *See also* black holes.

selection effect A term astronomers use for an effect that systematically distorts a measurement because it is impossible to obtain a fair sample. For example, if one tried to estimate the mean distance to a collection of stars, the measurement would be distorted by the fact that the brighter stars can be seen at greater distances. The objects are not selected correctly for a fair estimate.

shot noise The random fluctuations in light that come from the fact that it is composed of discrete photons rather than a continuous wave of energy. *See also* photon.

singularity A place where the equations of general relativity fail to predict the future. If a particle encounters a singularity, it has no future. Most singularities in known solutions are places of infinite curvature, which would tear apart any particle, but in principle they could be weaker. Physicists regard singularities as unsatisfactory, indicating that general relativity is incomplete. Most hope that quantizing gravity will eliminate them. The most serious singularities in known solutions are inside black holes, hidden from view, unable to influence us. The cosmic censorship hypothesis expresses the hope that this is generally true. *See also* curvature, quantum gravity, cosmic censorship hypothesis.

solar constant The mean flux of light from the Sun on the Earth. *See also* flux.

Solar Neutrino Unit A measure of the flux of neutrinos from the Sun at the Earth. One SNU is defined to be the flux that would induce one nuclear reaction in every 10^{36} chlorine atoms in a neutrino detector per second. *See also* flux, neutrino.

spacelike The term used to describe spacetime-intervals that are positive, so that the two events could be simultaneous in some frame. *See also* spacetime-interval, event, frame.

spacetime The set of all events. *See also* events.

spacetime diagram The graphical representation of events in spacetime. It shows a vertical axis which is the time as measured by some particular experimenter, and one or more horizontal axes for the space coordinates of events. Points in the diagram are events in spacetime. *See also* spacetime, events.

spacetime foam The idea that, through quantum effects, spacetime on a very small length-scale (the Planck length) could consist of constant fluctuations producing Planck-mass black holes. *See also* spacetime, quantum theory, Planck length, Planck mass, black holes.

spacetime-interval The invariant measure of distance or time in spacetime. The interval between two events is independent of the experimenter who measures time and space separations. It is the generalization of the Pythagorean theorem to spacetime. It is calculated from the metric tensor of spacetime, and it encodes the curvature of the spacetime. *See also* invariance, timelike, spacelike, Pythagorean theorem, curvature, metric tensor.

spacetime metric The metric tensor of spacetime, carrying the information about the spacetime-interval. *See also* spacetime-interval, metric tensor.

special relativity Einstein's first theory of spacetime, which treats how measurements of length and time are made. It predicts the slowing of time and the shortening of distances with motion. *See also* spacetime, time dilation, Lorentz–Fitzgerald contraction.

spectral lines Narrow features in the spectrum of light from an object, which indicate the presence of particular elements in the object.

spectroscopic binary A binary star system in which only one object is observed, and whose orbital motion is inferred from the time-dependent Doppler shift it produces in spectral lines from the observed star. *See also* spectral lines.

spin In fundamental physics, the intrinsic angular momentum carried by a particle.

spiral galaxy A galaxy which presents a spiral pattern in a photograph. Normally the spiral has two arms and it is tightly wound. The spirals are density waves, locations of rapid star formation, bright because the bright massive stars live for much less than the time it takes for the wave to move around the galaxy. *See also* elliptical galaxy, density wave.

spontaneous symmetry breaking In fundamental physics, the idea that a unified theory of forces can produce very unsymmetrical physical effects depending on random details of how the symmetries behave when the gas cools. *See also* electroweak, unified field theories.

standard candle The term astronomers use for an object whose intrinsic luminosity is known, so that its distance can be inferred from its apparent brightness. Most distance measurements in astronomy are calibrated by the use of standard candles.

standard meter stick By analogy with a standard candle, an object whose physical size is known, so that its distance can be inferred from its angular diameter. *See also* standard candle.

star The basic producer of light in the Universe. The Sun is a star. An object that is too small to produce light is a planet.

star cluster A set of stars grouped together and presumably formed together. Some clusters are large spherical assemblies of millions of stars; these are called globular clusters. Other clusters have fewer members more loosely bound to one another, and are called open clusters. *See also* globular clusters.

stationary limit The surface that is the outer boundary of the region around a rotating black hole (or possibly a very compact rotating star) in which there are orbits that have negative total energy with respect to a distant experimenter. This is associated with the fact that, inside the stationary limit, all free particles must rotate with respect to the distant experimenter: to stand still requires going faster than light with respect to local inertial frames. The negative-energy orbits have negative angular momentum relative to the hole. Non-rotating, Schwarzschild black holes do not have stationary limit surfaces. Kerr black holes have stationary limits that are topological spheres, called *ergospheres*. *See also* dragging of inertial frames, energy.

statistical mechanics The branch of physics that derives the macroscopic properties of gases from statistical averages over the motions and interactions of huge numbers of atoms and molecules. *See also* macroscopic, atoms, molecules, thermodynamics.

Steady-State model of the Universe The model devised by Hoyle and collaborators to show how a cosmology can be expanding and yet the same for all time. This requires matter to be created from nothing all over the Universe, to fill in the gaps left by the expansion of previously created matter. The hypothetical C-field was invoked to power this creation. Although the model permitted the Copernican principle to be applied to cosmology in its fullest sense, many physicists did not like the *ad hoc* nature of the C-field, and the model never gained many adherents. It has been modified substantially in recent years in order to be compatible with the cosmic microwave background radiation and other cosmological observations. *See also* C-field, Copernican principle, cosmic microwave background radiation.

strain The relative stretching or expansion of a system. For one-dimensional systems like rubber bands, this is defined as the change in length divided by the original length. In more than one dimension one must take into account shear as well as stretching. *See also* stress.

strange matter The whimsical name given to baryons that include the “strange quark” in their composition. Strange baryons can be made abundantly in accelerators. Some scientists speculate that strange matter is the real ground state of matter, so that normal matter will, in the right circumstances, transform itself into strange matter spontaneously. In this case, neutron stars could turn out to be strange stars. *See also* baryon, strange stars.

strange stars Stars made of strange matter. If strange matter is more stable than normal neutron matter, then these neutron stars might transform themselves into strange stars, which would be even more compact. *See also* neutron stars, strange matter.

stress The pressure, ram pressure, and shear forces inside smooth matter. Stress is what maintains strain in systems. *See also* strain.

stress–energy tensor The source of gravity in general relativity. Einstein knew he could not use just the mass density as the source, so he used the tensor that contains the mass-density, momentum-density, and pressure: the stress-energy tensor. *See also* strain.

string theory A candidate for the ultimate theory of all the physical forces. It describes particles as small loops of string, which is easier to treat consistently than a point-particle model. *See also* point mass, quantum gravity.

strong interaction The nuclear force, which is attractive enough to beat the electric repulsion of the protons in a nucleus and bind them together. Stronger than electric forces over distances the size of a proton, the strong interaction is a short-range force, falling off with distance much more rapidly than the electric force. So over the large size of an atom, it exerts a negligible influence. *See also* electromagnetism, nucleus, nucleon.

sub-millimeter A range of wavelengths of the electromagnetic spectrum with wavelengths smaller than 1 mm. *See also* infrared, microwave, X-ray, gamma-rays, ultraviolet radiation.

sunspots Dark blotches on the face of the Sun. These are cooler regions where magnetic fields loop out of the Sun and back into it.

superclusters Clusters of clusters of galaxies. These are the largest organized structures observed in the Universe. *See also* galaxy cluster.

superconductor A material that conducts electricity without resistance. This can only happen because of quantum effects. *See also* quantum theory.

superfluid A fluid that moves without friction. This can only happen because of quantum effects. *See also* quantum theory.

supergiant A massive giant star, very likely to explode as a supernova and leave behind a black hole or neutron star. *See also* supernova, black holes, neutron star.

supernova An explosion in a star that results in the destruction or dramatic transformation of the star. *See also* supernova of Type Ia, supernova of Type II.

supernova of Type Ia A supernova explosion which originates in a white dwarf and results in the complete disintegration of the star through a gigantic nuclear chain reaction. These supernovae are thought to be standard candles, and measurements of very distant examples have revealed the acceleration of the expansion of the Universe. *See also* white dwarf, supernova, standard candle.

supernova of Type II A supernova explosion which originates in the gravitational collapse of the inner core of a giant star, leaving behind either a neutron star or a black hole. *See also* gravitational collapse, supernova, giant, supergiant, neutron star, black holes.

surface of last scattering In cosmology, the location where photons became able to move freely through the Universe. *See also* decoupling.

surface brightness The apparent brightness of an object, per unit angular area on an image, i.e. per square radian on the sky.

synchronous rotation Rotation of a body about its axis with the same period as it executes an orbital motion, so that it always presents the same face to the object about which it orbits. The Moon is in synchronous rotation about the Earth.

tachyon A hypothetical particle that travels faster than light. Relativity forbids the acceleration of a particle up to the speed of light, but it is not inconsistent with relativity alone for a particle to travel faster than light, provided it never slows down to the speed of light. However, tachyons are difficult to reconcile with causality, since they travel backwards in time with respect to some observer.

tension *See* negative pressure.

tensor The mathematical generalization of the idea of a vector to something that effectively involves several vectors at once. A vector's components have one index. The next-simplest tensor is a matrix, whose components have two indices. It is possible to define tensors that have three or more indices. The full (Riemann) curvature tensor, for example, has four. *See also* matrix, vector, curvature, indices.

terrestrial planets The planets with rocky surfaces: Mercury, Venus, Earth, and Mars.

thermodynamics The study of heat, its transfer from one body to another, and the work that it can do in engines.

tidal acceleration The difference in the acceleration of gravity across an object or between two nearby objects. This is their relative acceleration, and it exists even in a freely-falling frame where their common acceleration is not present. Einstein identified the tidal acceleration as the true, observer-independent, non-removable effect of gravity.

time dilation The slowing of time produced by the motion of a body, as explained by special relativity. The faster a body's speed, the slower time goes. For photons, time stands completely still.

time travel Moving backwards in time relative to other objects, while time moves forwards for oneself. Physicists have discovered that wormholes can be used for time travel if they can be kept open for long enough for objects to pass through them, but keeping them open may require negative energy. *See also* wormhole, energy.

timelike The term used to describe spacetime-intervals that are negative, so that the two events could be at the same location in some frame. *See also* spacetime-interval, events, frame.

transponder An amplifier that receives a signal from a distant transmitter and returns the signal, amplified so that it has enough power to be received by the original transmitter. This allows the transmitter to measure the distance to the transponder from the round-trip travel time of the signal. This is the main way in which spacecraft are tracked from Earth as they move through the Solar System. A transponder can be thought of as an *active mirror*.

transverse Across the line of motion. The action of a water wave is transverse: the wave moves along the surface of the water, but the water itself moves up and down. *See also* polarization, longitudinal.

ultraviolet radiation Region of the electromagnetic spectrum with wavelengths shorter than the violet end of the visible spectrum, running from about $0.4\text{ }\mu\text{m}$ to 10^{-8} m . *See also* infrared, microwave, X-ray, gamma-rays, sub-millimeter.

unified field theories Ever since Maxwell unified the electric and magnetic forces into the theory of electromagnetism, in the nineteenth century, physicists have followed a path of simplifying the laws of physics by finding ways in which different forces are related. The electromagnetic and weak interactions have now been unified into the electroweak theory, which is very successful at explaining and predicting phenomena in its domain. Many physicists now expect that the next step will be a unification of the strong interaction with the electroweak. This would be a grand unified theory. *See also* electroweak, electromagnetism, weak interaction, strong interaction, grand unified theories.

vector The mathematical term for a directed object with length. *See also* component, displacement, position vector, indices.

virial method A method astronomers use to estimate the mass of a system of stars or galaxies by measuring their velocities relative to one another, and assuming that the mass creates a strong enough gravitational field to hold the objects together with roughly the same distribution of speeds over long periods of time.

viscosity Friction in the motion of a fluid.

visual binaries A binary system in which both stars are visible as separated images. Only the binaries nearest the Earth can be resolved in this manner.

visual magnitude The magnitude (brightness) of a star in visible light. *See also* absolute magnitude, apparent magnitude.

vortices A vortex (plural form: vortices) is the center of a rotational fluid motion. In everyday use it can be the center of a whirlpool, hurricane, or other natural phenomenon. In the theory of superfluidity, it is a line about which the fluid can rotate without having a rotational flow. This apparent contradiction requires a careful understanding of rotational motion in fluids. A superfluid is irrotational because a little stick or flag embedded in the fluid will not spin as the fluid moves. Nevertheless, the fluid can flow past obstacles on curved paths, provided it always keeps locally irrotational. It can even rotate about a center if

its angular velocity about the center is proportional to $1/r$, the distance from the center. This implies that the fluid description breaks down at the center, and so the vortex at the center is a kind of singular point, not a place where the fluid remains superfluid. Related phenomena allow magnetic fields to thread through superconductors, and also lead to cosmic strings in the expanding Universe. *See also* cosmic strings.

watt The unit used in the SI system for power, abbreviated W: $1\text{ W} = 1\text{ kg m}^2\text{ s}^{-2}$. *See also* power.

weak interaction The force between leptons. It is the force that is responsible for beta decay. For example, the decay of a neutron into a proton, electron, and electron neutrino is a beta decay, because it requires the creation of two leptons. *See also* beta decay, strong interaction.

weight The force of gravity on an object, proportional to its mass but also to the acceleration of gravity.

white dwarf A star whose support against gravity comes from the pressure of a degenerate gas of electrons. Its typical mass is about a solar mass, and its radius about that of the Earth. *See also* degenerate gas, neutron star.

white hole The time-reverse of a black hole, in which a compact object exists from the distant past, but nothing can fall into it; instead, it explodes and disappears. This behavior is possible according to general relativity, but would require very special initial conditions at the Big Bang to make the initial compact objects. There is no evidence for such objects in the real Universe. *See also* black holes.

work The product of force and distance, which represents the energy expended by the force in moving an object through the

given distance. Work is measured in energy units, like joules. *See also* joule.

world line The set of events experienced by an object over time. This is a timelike line through spacetime.

wormhole A hypothetical connection between one region of space and another, or even between our Universe and another. They are relatives of black holes. Although no such connections have ever been observed, they are interesting theoretical objects because they allow physicists to explore the limits of what might be possible in general relativity. Even time travel may be possible with such objects. *See also* black holes, time travel.

X-ray Region of the electromagnetic spectrum with shorter wavelengths than ultraviolet. Typical wavelength range is 10^{-11} m to 10^{-8} m . For X-rays and gamma-rays it is typical to quote energies rather than wavelengths; this range runs from about 0.1 keV to 100 keV. *See also* infrared, ultraviolet radiation, microwave, gamma-rays, sub-millimeter.

X-ray binary A binary star system that emits X-rays because gas is flowing from one of the stars (usually a giant) onto its companion (usually a neutron star or black hole). The energy released as gas spirals down onto the companion heats the gas to temperatures where it emits X-rays. *See also* X-ray, giant, neutron star, black holes.

zero-point energy In quantum theory, nothing can be perfectly at rest, so when a material is cooled to absolute zero, its atoms retain a small motion, whose energy is the zero-point energy. *See also* quantum theory.

Index

Page numbers in *italics* refer to terms in figures and investigations.

- 0538-641, 299
1956+350, 299
3C273, 175, 176
4U1543-47, 299
- A0620-00, 299
absolute magnitude, *see* magnitude, absolute
absolute temperature, *see* temperature, absolute
absolute velocity, *see* velocity, absolute
absolute zero, *see* temperature, absolute zero
abundance of elements, *see* elements, abundance
acceleration, 2–5, 9–12, 14, 15, 18, 19, 21–23, 25–35, 38, 40, 41, 43, 52, 53, 73, 74, 78, 91, 92, 118, 156, 157, 168, 169, 190, 193, 194, 196, 200, 205, 221, 222, 243, 247–250, 253, 268, 273, 280, 291, 303, 305, 313, 320, 359, 360, 362, 363, 366, 371, 388, 400, 404, 421, 426, 429, 432, 433, 437, 439, 440
average, 3
circular, *see* circular motion, acceleration
instantaneous, 3, 22, 169
magnetic-type, 248, 250
of gravity, *see* gravity, acceleration of
accelerator, particle physics, 328
accretion, 159–161, 278, 299, 300, 302, 350, 421, 433, 435, 437
disk, 158–161, 177, 278, 299–302, 421, 431
center, 159, 160
action and reaction, *see* Newton, laws of motion, third
action at a distance, 13, 14, 310, 421
active curvature mass, 240, 244, 245, 315, 405, 421
active galaxy, 167, 299, 347, *see also* quasar nucleus, 176
active gravitational mass, 240, 242–247, 250, 253–255, 280, 315, 351, 358, 360–362, 378, 388, 398, 400, 405, 421, 426, 429, 434
ad hoc, 254, 395, 439
Adams, John C, 48
AEI, *see* Max Planck Institute for Gravitational Physics
AIGO, *see* interferometric detector, AIGO
airplane, 4, 6, 9, 20, 21, 72–74
alpha particle, 122, 132, 133
Amazon river, 39
American Institute of Physics (AIP), 135, 269
Andromeda galaxy, 47, 104, 170, 363, 434
anaerobic bacteria, 69
angle of inclination, 156, 421, 433
Ångstrom, 301
Ångstrom, 87, 301, 302, 421
angular momentum, 32, 45, 52, 58, 64, 113, 142, 268, 276, 291, 296–299, 307, 421, 423, 424, 428, 431, 434, 438, 439
conservation, *see* conservation law, angular momentum
anisotropy, 357, 384, 421
Anthropic Principle, 134, 372, 401, 413, 421
anti-gravity, 244, 259, 351, 358, 421, 425, 426
anti-matter, 368, 369, 375, 412, 416, 421
anti-neutron, 368, 369, 375, 412
anti-proton, 368–370, 375, 412, 422
aphelion, 58, 422
APM, *see* Automatic Plate Measuring machine
Apollo project, 65, 126, 422
apparent brightness, *see* brightness, apparent
apparent magnitude, *see* magnitude, apparent
Archimedes, 72
argon, 128
Arecibo radio telescope, 317
Aristotle, 255
arrow of time, 307, 415, 416, 422
Association of Universities for Research in Astronomy (AURA), 176
asteroid, 61, 62, 90, 91, 159, 422, 432, 436
belt, 46, 159
astronaut, 2, 19–21, 54, 65, 126, 188, 189, 221, 303, 422
astronomical unit, 27, 28, 51, 56, 60, 61, 103, 104, 106, 162, 176, 339, 343, 419
astronomical unit (AU), 27, 422
atmosphere, 19, 65–73, 75, 77, 79–83, 89, 91, 94, 98, 105, 114, 159, 198, 199, 201, 205, 300, 378, 406, 421, 425
Earth, 19, 65, 67–71, 79–81, 84, 93, 105, 135, 153, 199, 210, 435
Mars, 68
planetary, 80, 94, 95
Venus, 68, 69, 83
atmospheric pressure, *see* pressure, atmosphere
atom, 65, 69–71, 76–79, 81, 84, 85, 87, 89, 90, 97, 100, 103, 112–116, 118–124, 128, 143, 153, 171, 173, 181, 196, 198, 200, 252, 257, 261, 262, 280, 301, 306, 328, 354, 363, 365, 366, 373, 375, 377, 395, 406, 410, 412, 421–424, 427, 428, 431–439, 441
atomic clock, *see* clock, atomic
atomic weapon, *see* nuclear bomb
atomic weight, 81
AU, *see* astronomical unit
AURA, *see* Association of Universities for Research in Astronomy
AURIGA, *see* bar detector, AURIGA
aurora, 85, 98, 272
aurora australis, 98
aurora borealis, 98
Automatic Plate Measuring (APM) machine, 379
axis of rotation, *see* rotation, axis of
- B1359+154, 331
bar detector, 326–328, 422, *see also* gravitational wave detector

- AURIGA, 327
 spherical, 327
- baryon, 127, 222, 296, 378, 381, 422, 437, 439
 baryonic charge, 127
- basketball, 263, 264
- Bay of Fundy, 44
- beam splitter, 187
- beaming, relativistic, 301, 422
- Bekenstein, J, 307
- Bell, J, 271, 272, 319, 355
- Bernoulli effect, 74
- Bernoulli, D, 73, 76
- beta decay, 127, 129, 411, 412, 422, 427, 434, 441
- bicarbonate, 67
- Big Bang, 6, 77, 121, 124, 126, 131, 132, 134, 174, 175, 179, 210, 212, 260, 275, 293, 295, 325–327, 330, 345–347, 350–355, 357, 358, 365–367, 369, 370, 372–376, 378, 381, 391, 393, 394, 398, 401, 408, 413, 415, 422, 428, 430, 441
 before, 346, 416
 black hole formation, 293
 repeating, 134
- Big Crunch, 134, 358, 364, 413, 422
- Big Freeze, 358, 422
- binary stars, 23, 29, 46, 90, 151, 153–155, 157, 160–162, 165, 239, 259, 265, 269, 276–279, 292, 299, 309, 310, 318–322, 324, 350, 421, 423, 424, 429, 433, 438, 440, 441
 breakup, 269
 chirping, 322, 423
 as standard candles, 321
 characteristic time, 319, 321, 423
 distance to, 321
 gravitational wave signal, 321, 322, 423
 close, 160, 162, 165
 coalescing, 320, 322, 325
 orbit, 154–157, 269, 276, 433
 spectroscopic, 154, 156, 438
 visual, 153, 156, 440
- binding energy, 87, 165, 171, 264, 265, 412
 cluster, 165
- binocular vision, 333
- binomial theorem, 43, 192
- birds, aerodynamics, 6, 12, 74, 222
- BL lac objects, 176
- black body, 110–118, 136, 137, 161, 304, 305, 354, 356, 375, 408, 422
 luminosity, 117, 265, 267
 model of Sun, 118
 radiation, 109, 113, 267, 272, 277, 278, 304, 305, 356, 394, 422, 430
 spectrum, 112, 114, 116, 117, 300, 304, 354–357, 422, 425, 430
- black hole, 2, 6, 11, 23, 34, 36–38, 84, 103, 120, 134, 135, 140, 143, 149–151, 153, 160–163, 164, 165, 167, 168, 171, 172, 174–177, 179–181, 187, 193, 207, 210, 212, 219, 221, 231, 238–241, 243, 251, 252, 258, 260, 265–267, 267, 268, 278, 279, 282, 284–308, 310, 313, 318, 321–324, 339, 341, 346, 347, 350, 360, 381, 391, 394, 404, 407, 408, 411, 415, 416, 422–425, 427, 429–433, 435–441
 as gravitational lens, 341
 binary, 162, 165, 322
 coalescence, 239, 279, 292, 322–324
 center, 408
 collision, 323, 330
 density at formation, 292, 293, 322
 Einstein's views on, 285
 evaporation, 415
 horizon, 287–293, 296–300, 302, 304, 305, 307, 308, 322, 323, 341, 346, 347, 359, 389, 408, 422, 425, 430, 434, 435, 438
 interior, 293, 417
 Kerr, 177, 239, 296–298, 439
 ergosphere, 297–299, 427, 435, 439
 lifetime, 305
 mass, 286, 292, 299, 305, 308, 438
 massive, 164, 166, 175, 300, 301, 309, 322, 381
 model for quasar, 343
 Newtonian version, 36, 37, 287
 primordial, 305, 436
 rotating, *see* black hole, Kerr Schwarzschild
 geometry, 286–288, 291, 292, 295–298, 302, 303, 305, 320, 431, 438
 radius, 286–288, 290, 291, 293, 294, 411, 438
 small mass, 293
 supermassive, 161, 293, 299, 345
 uniqueness of, 296
- black holes
 merger, 167
 blueshift, 15, 16, 192, 422
- Bohr, N, 84, 116, 395
- Boltzmann's constant, 76–78, 89, 111, 300, 419
- Boltzmann, L, 69, 70, 76–79, 83, 84, 88–91, 111, 112, 116, 254, 300, 306–308, 419
- Boomerang, *see* cosmic microwave background (CMB), Boomerang experiment
- boron, 127, 129
- Bose, S N, 145
- boson, 143, 145, 423, 428, 432, *see also* fermion
 boson star, *see* star, boson
- bounce, 76, 87, 101, 148, 149, 151, 266, 283, 432
- brane, 409, 423
- bremsstrahlung, 133, 423
- brightness
 apparent, 105, 106, 108, 168, 316, 422, 431, 433, 438, 440, *see also* magnitude, apparent
 intrinsic, 108, 168, 433, *see also* magnitude, absolute
- Bristol Channel, 44
- brown dwarf, 172, 338, 423
- Brownian motion, 88, 423
- buoyancy, 71–75, 102, 309, 423
 neutral, 72
- C-field, 124, 125, 423, 439
- calcium, 67, 121
 calcium carbonate, 67
 cycle, 121
- calculus, 4, 9, 13, 28, 34, 73, 111, 119, 161, 232, 258, 423, 426, 428
 and finite differences, 4
- Cambridge University, 271
- cannonball, 2–6, 19, 20, 22, 23
- carbon, 66, 68, 121, 125, 127, 132, 133, 135, 139, 143, 147, 148, 151, 210, 302, 370, 411, 426, 435
 burning, 140
 carbon dioxide, 66–69, 430
 cycle, 67, 68, 121, 127
 excited state of nucleus, 132, 133, 428, 432
 nuclear energy levels, 411
- carbonic acid, 67
- Cartwheel Galaxy, 166
- Casimir effect, 396, 398

- Casimir, H, 396
cataclysmic variable, *see* variable stars, cataclysmic
catalyst, 127, 423
causality, 196, 203, 216, 414, 415, 417, 440
cause and effect, *see* causality
caustic, 336, 341
Cavendish, H, 31, 36, 37
CCD, *see* charge coupled device
CDM, *see* cold dark matter (CDM)
celsius, 75, 423, 432
center of mass, 156
centigrade, 423
centimeter-gram-second (cgs) system of units, 419
centrifugal effect, 98, 423
centrifuge, 21
Cepheid variable, *see* variable stars, Cepheid
CERN, 374
cgs, *see* centimeter-gram-second system of units
Chandrasekhar mass, 145–148, 151, 160, 207, 264–266, 277, 286, 350, 410, 411
Chandrasekhar, S, 145, 146, 267, 286, 316
characteristic frequency, *see* frequency, characteristic
characteristic vibration, *see* mode
charge
 electric, 124, 127, 128, 133, 142, 172, 206, 248, 296, 297, 311, 398, 421, 424, 427, 436
 magnetic, *see* magnetic monopole
charge coupled device (CCD), 436
Charles' law, 75, 76, 78
Charles, J, 75, 168, 254
chemical bond, 87
chemical element, *see* element, chemical
chemical reaction, 82, 123, 124, 159, 423
chemistry, 69, 70, 82, 93, 123, 134, 135, 263, 410–412, 431, 434
China, 44, 274
chirp, *see* binary stars, chirping, gravitational wave signal
chirp time, *see* binary stars, chirping, characteristic time
chlorine, 128, 130, 210, 438
circular motion, 22, 26, 28, 40, 41, 156, 169, 320, 324, 423, 424, 426
 acceleration, 21, 33, 156
- circular orbit, *see* orbital motion, circular
circumference, 19, 22, 26, 35, 100, 280, 287, 304, 305, 348, 385–387, 389, 390
circumferential radius, *see* radius, circumferential
CLOO24+1654, 332
classical (non-quantum) physics, 295, 307, 308, 315, 407, 408, 410, 414, 424, 426, 427, 430, 437
Clausius, R, 306, 307
clock, 3, 17, 131, 188, 191, 195, 197–200, 201–203, 208, 215, 221, 230, 231, 233, 272, 277, 286–288, 291, 318, 389, 400, 437
 atomic, 17, 275
 synchronize, 197, 199, 200, 201, 203, 204, 208, 216, 428
CMB, *see* cosmic microwave background (CMB)
co-latitude, 228, 424
coalescing binary, *see* binary stars, coalescing
COBE satellite, *see* cosmic microwave background (CMB), COBE satellite
cold dark matter, 173
cold dark matter (CDM), 173, 404, *see also* dark matter
cold dark matter (CDM), 366, 377, 380, 424, 425
cold dark matter (CDM), 378–381
color
 and temperature, 85
 index, 110, 424
 of a star, 15, 85, 87, 88, 109–113, 116, 118, 139, 140, 222, 296, 421, 424
 of light, 109
Colorado, 35
comet, 29, 30, 51, 61, 90, 91, 157, 435
component, 20–23, 35, 45, 78, 100, 231, 253, 291, 300, 319, 321, 382, 391, 421, 424, 426, 427, 431, 440
compose, 183, 424
composition, 65, 66, 70, 82, 93, 95, 99, 101, 110, 114, 131, 137, 139, 143, 145, 147, 148, 183, 192, 306, 371, 382, 421–424, 435, 439
 solar, 94, 129
- velocity, *see* velocity-composition rule
Compton gamma-ray satellite, 279
Compton scattering, 96, 133, 279, 424
Compton, A H, 96, 279
computer, 3–6, 22, 23, 28, 30, 34, 36, 44, 48, 55, 56, 70, 73, 81, 82, 92, 94, 97, 111, 119, 135, 138, 139, 148, 149, 154, 155, 157, 158, 162, 166, 240, 290, 292, 322, 323, 332, 359, 362–364, 367, 371, 373, 379, 427, 428
model, 70, 79, 82, 85, 92, 95, 96
program, 4–6, 19, 23, 24, 28–33, 35–38, 49, 55–57, 62, 65, 71, 73, 80–82, 91–97, 119, 136, 139, 154–157, 160, 162, 280, 282, 290, 291, 341, 356, 364, 433
conservation law, 54, 55, 127
 and invariance, 64
angular momentum, 32, 45, 52, 58, 64, 268, 424
energy, 54–59, 62, 63, 204, 205, 232–234, 256–258, 296, 298, 317, 335, 380, 424, 431, 436
momentum, 52, 54, 206
convection, 70, 82, 92, 94, 99, 135, 424
converge, 30, 95, 132, 220, 224, 261
coordinates, 21–23, 29, 35, 36, 157, 183, 213, 214, 217, 224, 226–229, 231, 232, 234, 235, 239, 241, 251, 256–259, 285, 287, 288, 291, 320, 355, 356, 389, 424, 426, 428, 431, 433, 434, 437, 438
 polar, 227, 229
 rectangular, 22, 23
 spherical, 227, 424
Copernican principle, 345, 347–349, 353–355, 357, 389, 393, 424, 437, 439
Copernicus, N, 6, 18, 25, 347, 424
correlation, 326
cosine rule, 229
cosmic background radiation, *see* cosmic microwave background (CMB)
cosmic censorship hypothesis, 290, 298, 425, 434, 438
cosmic gravitational wave background, *see* gravitational waves (in astronomy), cosmological
cosmic microwave background (CMB), 177, 325, 353–358, 360,

- 362, 364, 365, 367, 370, 375,
 378–381, 384, 391, 392, 402,
 406, 425, 439
 Boomerang experiment, 381
 COBE satellite, 163, 354, 356, 380,
 381
 Maxima experiment, 381
 Planck satellite, 391
 cosmic rays, 128, 135, 137, 198, 205,
 210, 382, 392, 406, 407, 425
 flux, 392
 highest-energy, 392, 406
 cosmic string, 173, 223, 244, 377,
 378, 381, 392, 405–407, 425,
 441
 no curvature of time, 405
 cosmological constant, 193, 244,
 253–259, 350, 351, 361–363,
 371, 379–381, 388, 392–400,
 402, 403, 407, 425, 437, *see also*
 dark energy
 density, 255
 dimensions, 254
 inertia, 255
 pressure, 254
 pressure invariance, 255
 cosmological fluid, *see* cosmological
 constant
 cosmological inflation, *see* inflation
 cosmological model, *see* cosmology
 cosmological redshift, *see* redshift,
 cosmological
 cosmological singularity, *see* singularity,
 cosmological
 cosmology, 63, 120, 127, 131, 134,
 151, 168, 177, 178, 180, 187,
 193, 238, 240, 241, 244, 254,
 260, 308, 342, 343, 345–347,
 351, 352, 357, 359–361, 363,
 364, 367, 370, 371, 374, 378,
 383, 384, 388–391, 393–395,
 399, 405, 414, 417, 425, 430,
 431, 434, 439, 440, *see also*
 Universe
 Big Bang, *see* Big Bang
 closed, 384, 387
 matter-dominated, 388
 critical density, 361–364, 373, 379–
 381, 388–390, 392, 397, 398,
 401–403, 413, 425, 426
 density parameter, 361, 362, 426
 expansion, 319, 363, 365, 366, 377
 exponential, 371, 401
 flat, 384, 385
 Friedmann, 385, 387, 389, 390, 393
 hyperbolic model, 388
 in general relativity, 383, 384
 in Newtonian gravity, 359, 360,
 383, 384
 matter-dominated, 360–362, 364,
 366, 371, 376, 388, 434
 model universe, 163, 254, 257, 260,
 303, 306, 307, 331, 348, 349,
 352, 353, 357, 358, 361–365,
 370, 383–385, 387–390, 392,
 393, 402, 417, 430, 434, 435
 nucleosynthesis, 358
 open, 384, 388
 matter-dominated, 388
 our Universe, *see* Universe
 physical, 367, 436
 radiation-dominated, 370, 371,
 376, 400
 rubber-band universe, 348, 349,
 352, 363, 365, 439
 scale-factor, 348, 363–365, 370,
 371, 375, 376, 400, 425
 standard model, 353
 steady-state, 125, 358, 423, 439
 visible density, 362
 Coulomb field, 243
 covariance principle, 239–241, 256,
 258, 259, 437
 Cowan, C, 123
 Crab Nebula, 150, 267, 270, 271
 pulsar, 272–275, 325
 Crab nebula
 supernova, 274
 creation of the Universe, 111, 125,
 127, 240, 256, 258, 345, 357,
 358, 398, 409, 432, 439, 441,
see also Big Bang
 curvature, 213, 218, 221, 223–225,
 229, 230, 232, 234–237, 239,
 240, 242, 244, 245, 256, 258,
 259, 280, 287, 289, 305, 310,
 315, 385, 388–390, 402, 405,
 421, 423, 425, 427–429, 436,
 438, 440
 describes tidal forces, 224
 gravitational, 224
 space, 14, 222, 225, 229, 235, 236,
 239, 240, 280, 287, 289, 312,
 361, 383, 405, 421, 429, 434
 spacetime, 222, 224, 225, 230, 231,
 245, 258, 260, 280, 281, 288,
 289, 304, 337, 390, 425, 427–
 429, 434
 time, 213, 225, 230, 232, 234, 236,
 239–243, 286, 288, 289, 310,
 312, 320, 378, 392, 402, 405,
 415, 421, 425, 427, 429
 cycles, 16, 25, 87, 121, 288, 315, 322,
 425, 427
 CYG X-1, 299, 300
 dark energy, 351, 353, 362, 389, 392,
 393, 398, 402–404, 413, 425,
 426, 434, 437, *see also* cosmo-
 logical constant
 dark matter, 105, 163, 170, 172, 173,
 177, 212, 331, 332, 338, 342,
 343, 351, 353, 360–362, 366,
 368, 377–382, 391, 392, 398,
 402, 404, 409, 412, 413, 424–
 426, 428–430, 436
 cold, *see* cold dark matter (CDM)
 dark clusters, 172
 particles, 173, 368, 378, 379, 382,
 404, 407, 412, 430
 dark side of the Universe, 324
 Darwin, *see* life, evolution of
 Davis, R, 128–131
 deceleration parameter, *see* Universe,
 deceleration
 deflection of light, 37, 38, 163, 212,
 225, 235–238, 245, 256, 260,
 261, 277, 285, 331, 332, 336,
 404, 405, *see also* gravitational
 lens
 by the Sun, 34, 225, 239
 in Newtonian gravity, 38
 radio waves, 318
 degenerate gas, 141, 280, 426, 435,
 441
 degenerate matter, 145
 electron gas, 142, 144, 146, 147
 Fermi, 142, 144, 146, 147
 neutron gas, 147, 148, 280
 pressure, 142, 144–148
 proton gas, 144, 146
 degenerate star, *see* star, degenerate
 demagogue, 208, 209
 density, 31, 35, 43, 44, 73, 75, 77–
 79, 81, 87, 89, 91–95, 98, 100,
 101, 105, 120, 124, 129, 137,
 138, 144, 145, 147–149, 151,
 159, 173, 180, 193, 194, 240–
 245, 247, 250, 253–258, 261–
 266, 272, 280–284, 289, 292,
 293, 295, 306, 311, 350, 353,
 356, 358, 360–364, 366, 370–
 374, 376–381, 388–393, 395,
 397, 398, 400–403, 405, 412,
 413, 421–423, 425, 426, 436–
 439

- irregularities in, 131, 177, 357, 368, 377–380, 394, 401, 402, 436
of water, 89, 144, 261, 293
- derivative, 161, 258, 426–428
- deuterium, 127, 130, 133, 372–374, 375, 426
- differential equation, 73, 232, 258, 426, 427
- Digitized Sky Survey, 104, 164, 165
- dimensions, 5, 113, 117, 142, 144, 146, 156, 195, 213, 214, 219, 221, 230, 249, 254, 256, 274, 294, 315, 348, 349, 356, 363, 385, 387, 388, 400, 408, 409, 416, 419, 423, 424, 426, 427, 431, 436
- dimensional analysis, 5, 113, 315, 400, 426
- dimensionless number, 99, 117, 183, 274, 321, 410, 426
- dirigible, 73
- disk, 66, 131, 158, 159, 161, 165, 166, 174, 176, 177, 252, 267, 270, 278, 300, 301, 328, 421, 433
- Disneyland, 213
- displacement, 22, 62, 311, 312, 328, 386, 426
- distance
proper, *see* proper distance; proper length
- DNA, 87
- Doppler effect, 15, 16, 154, 156, 160, 175, 182, 187, 192, 204, 276, 301, 317, 357, 364, 365, 384, 438
transverse, 192
- dragging of inertial frames, *see* Lense-Thirring effect
- Drake, F, 131, 132
- drum, 99, 100
- dust, 25, 66, 83, 88, 135–137, 166, 167, 178, 309, 341, 423, 426, 431
interstellar, 65, 66, 90, 135, 159, 166, 296, 426
- Dyson, F W, 38, 331
- Eagle Nebula, 136, *see* M16
- early warning, 324
- Earth, 3, 6, 10, 12, 14–26, 26, 27–29, 31, 33, 34, 36, 38–46, 48, 51, 52, 56–58, 60–62, 64–71, 77, 79–86, 90, 91, 93, 98, 99, 101, 103–108, 113, 114, 116, 118, 121–123, 125–131, 134, 135, 144, 145, 149, 151, 153, 156, 176, 177, 179, 184–188, 199, 201, 203, 205, 207–211, 213, 219, 222, 223, 227, 228, 230–232, 241, 251–253, 259, 261, 264, 266, 270–276, 281, 283, 284, 288, 309, 310, 312, 315, 317, 319, 320, 324, 325, 329, 330, 332, 345, 347, 354–356, 367, 371, 376, 387, 392, 399, 406, 412, 413, 417, 419, 421, 422, 424–426, 429, 430, 432, 433, 435–438, 440, 441
- age of, 122
- center, 21–23, 33, 34, 39, 41, 45, 252, 421
- crust, 46
- density, 31
- escape speed, *see* escape speed, Earth
- mass of, 27, 29, 31, 36, 153, 276, 419
- orbit, 43
- radius, 19, 22, 26, 27, 33, 35, 69, 79, 103, 116, 118, 143, 144, 146, 228, 419
- rocks, 126
- rotation rate, 45
- temperature, 67, 80, 93, 113, 430
- volcanic activity, 46
- Earthlings, 207
- eccentricity, 28, 29, 32, 238, 277, 319, 426
- eclipse, 38, 42, 43, 115, 235, 331, 426
solar, 38
- Eddington, A, 38, 145, 331
- Einstein
curvature tensor, 258, 427
field equations of general relativity, 236, 238–240, 242, 244, 245, 248, 253, 256, 258–260, 281, 285, 292, 295, 296, 313, 316, 367, 388, 407, 425, 427
- luminosity, 294, 305, 320, 427
- radius, 336–338, 342, 427
- ring, 335, 337–339, 427
- Einstein Cross, 332, 337, 342, 343
- Einstein velocity-composition rule, 183, 187
- Einstein, A, 6, 9, 11, 15, 17, 21, 36, 38, 40, 85–89, 112, 113, 116, 122, 178–184, 184, 185, 186, 189, 190, 195, 197, 199, 201, 206, 211–214, 216, 219, 222–225, 232, 234, 236–242, 244–246, 250, 252–259, 285, 286, 309, 310, 313, 316, 318, 337, 351, 359, 383, 384, 394–397, 402, 415–417, 425, 427, 432, 436–440
- as a superstar, 236
- greatest blunder, 256
- electromagnetic waves, 17, 85, 87, 112, 187, 268, 273, 322, 324, 325, 354, 355, 395, 423, 434, 436, *see also* light; radio waves
- transversality, 311
- electromagnetism, 12, 15, 180, 182, 212, 245, 248, 268, 304, 311, 365, 391, 395, 398, 408, 421, 427, 440
- electric field, 243
- Maxwell's equations, 182
- electron, 12, 70, 81, 86–89, 91, 96, 97, 114, 118–120, 122–124, 127, 130, 131, 133, 142–148, 172, 173, 177, 180, 195–197, 204, 205, 207, 210, 260, 262–264, 267, 268, 279, 301, 305, 311, 346, 354, 365, 368–370, 372, 374, 375, 375, 377, 378, 382, 391, 402, 408, 410–412, 419, 422, 425, 427, 428, 431, 432, 434–436, 441
- mass, 124, 131, 133, 143, 348, 410, 419
- neutrino, *see* neutrino, electron-type
- electron volt (eV), 88, 89, 119, 120, 123, 124, 159, 205, 367, 375, 376, 382, 392, 406, 419, 427, 430
- electrostatic, *see* force, electric
- electroweak interaction, *see* force, electroweak
- element
chemical, 103, 135, 423
- elementary particle, *see* particle, elementary
- elements
abundance, 165
ages of, 125
heavy, 65, 66, 69, 94, 125–127, 131, 132, 151, 177, 210, 261, 266, 301, 349, 367, 370, 411, 412, 426, 433
- elevator, 21
- ellipse, 13, 22, 25, 27, 28, 32, 48, 58, 154–157, 169, 212, 237, 269, 276, 310–312, 312, 318, 320, 426, 436
- focus, 13, 28

- ellipsoid
oblate, 98
prolate, 98
- elliptical galaxy, 164, 165, 166, 168, 170, 174, 176, 427, *see also* galaxy
- giant, 166, 167, 170, 176
formation, 167
- elsewhere, *see* spacetime structure, elsewhere
- energy, 53–59, 61–63, 67–69, 76, 82, 84–91, 93, 95, 96, 99, 103, 105–107, 109, 111–117, 119–125, 127–129, 133–135, 137–142, 144–149, 151, 157–162, 164, 165, 171, 174–179, 187, 190–195, 200, 201, 204–207, 210, 232–234, 240, 241, 244, 254, 256–258, 264, 265, 267, 273–275, 277–279, 281, 284, 292, 294–296, 298–301, 303–308, 310, 313, 315–317, 319–321, 324, 325, 335, 350, 351, 353, 354, 356, 361, 362, 367–372, 375, 375, 376–380, 382, 389, 391–393, 395, 396, 398–407, 409–413, 417, 424–441
- conservation, *see* conservation law, energy
- density, 241, 254, 255, 257, 356, 361, 370, 371, 374, 397, 400, 405, *see also* mass, density; mass-energy, density
- duty, 123, 124, 141
- electrical, 54
- flux, 106, 111, 117, 314–317, 320, 356, 422, 428
- gravitational, 54, 90, 122, 123, 137, 159, 161, 162, 171, 172, 240, 273, 275, 278–280, 298, 303, 316, 423, 437
- gravitational wave, *see* gravitational waves (theory), energy
- inertia of, 190
- kinetic, 54, 55, 57, 59, 62, 69, 70, 76–78, 80, 84–86, 88, 90, 91, 123, 127, 137, 143–145, 149, 151, 159, 161, 166, 171, 175, 190, 192–194, 205, 206, 232, 233, 250, 257, 264, 265, 273, 275, 298, 301, 303, 316, 320, 399, 432, 436
- of the vacuum, 400, 401, 405
- orbital, 53, 159, 321
- potential, 54, 55, 57, 62, 63, 90, 137, 161, 171, 172, 240, 273,
- 275, 280, 298, 303, 316, 321, 423
relation to time, 63, 87, 204
relativistic, 190, 233
zero-point, 84, 395–397, 399, 441
- entropy, 133, 296, 306–308, 422, 427
- Eötvös, R, 11
- epicycles, *see* planets, orbits, epicycles
- equation of state, 77, 78, 92, 93, 95, 101, 144, 146, 257, 280–283, 431
- equivalence principle, 2, 10–18, 20, 21, 27, 36–40, 74, 75, 211, 222, 224, 225, 231, 236, 241, 258, 260, 266, 313, 317, 318, 331
and geodesics, 224
- ergosphere, *see* black hole, Kerr, ergosphere
- ESA, *see* European Space Agency
- escape speed, 35, 36, 51–53, 55, 56, 70, 86, 100, 123, 137, 147, 148, 160, 162, 165, 175, 179, 180, 264–266, 269, 358, 361, 363, 388
- Earth, 36, 52
- equivalent energy, 161, 171, 234, 264, 265
- Sun, 37, 51, 53, 60
- escape velocity, *see* escape speed
- ether, 182, 185, 199, 255
friction with, 185
- Euclidean geometry, 219, 226, 427
- Euclidean plane, 213, 385, 427, 432
- Euclidean space, 223, 229, 234, 235, 280, 287, 292, 385–389, 427–429, 437
- European Space Agency (ESA), 19, 85, 98, 102, 105, 140, 309, 329, 391
- event, *see* spacetime structure, event
- event horizon, *see* black hole, horizon
- exclusion principle, *see* Pauli exclusion principle
- experimenter, *see also* relativity, experimenters or observers
- freely-falling, 15, 16, 18, 20, 21, 38–41, 148, 224, 232, 241, 251, 258, 266, 280, 281, 288, 293, 297, 303, 312, 426, 429, 440, *see also* equivalence principle
- exponential, 119, 371, 393, 400, 402, 426, 433
- exponential function, 111, 119, 371, 428
- Fermi sea, *see* degenerate gas, Fermi
- Fermi, E, 142, 144, 145, 205, 262
- fermion, 143, 145, 423, 426, 428, *see also* boson
- Feynman, R P, 245, 246, 288, 395
- filter, 110
blue, 110
infrared, 110
ultraviolet radiation, 110
visual, 110
- finite differences, 4, 29, 85, *see also* computer model; computer program; calculus, and finite differences
- Fitzgerald, G F, 185, 186, 188, 199
- flat space, *see* Euclidean space
- flux, 106–108, 111, 117, 128–131, 314–317, 320, 356, 392, 422, 424, 425, 428, 438, *see also* apparent luminosity; energy, flux; cosmic rays, flux; solar neutrinos, flux
- neutrino, *see* solar neutrinos, flux per unit wavelength, 111
- force
- electric, 22, 63, 213, 246, 263, 439
- electromagnetic, 127, 243, 263, 411, 427, 432
- electroweak, 127, 427, 429, 440
- fundamental forces, 127, 346
- gravitational, 13, 15, 20, 27, 31, 33, 35, 40, 41, 43, 45, 46, 79, 103, 157, 159, 160, 167, 170, 232, 246, 247, 257, 289, 329, 359, 361, 431
- spherical, 34
- nuclear, 63, 128, 131, 133, 263, 411, 439
- hard-core, 263, 265, 266, 411, 412
- hard-core repulsion, 282, 293, 411
- range, 263
- strong, 85, 127, 427, 429, 431, 435, 439, 440
- weak, 127, 128, 264, 427, 429, 441
- frame, *see* relativity, experimenters or observers
- frequency, 15–17, 87, 99–102, 112–115, 117, 118, 162, 182, 191, 192, 197, 204, 232, 254, 273, 276, 288, 293, 301, 314–317, 319–322, 325, 328–330, 370, 396, 414, 423, 428, 430, 432, 435, 436

- characteristic, 99–101, 423, 428, 430, 435
- fundamental, *see* mode, fundamental, frequency
- friction, 9, 12, 44, 46, 62, 111, 122, 148, 156, 157, 159, 161, 278, 284, 310, 318, 435, 439, 440
- Friedmann cosmology, *see* cosmology, Friedmann
- Friedmann, A, 219, 384, 388
- fundamental physics, 120, 127, 131, 132, 173, 182, 183, 256, 261, 391, 392, 401, 404, 407, 412, 426, 438
- fusion reactor, 138, 354
- G*, *see* gravitational constant *G*
- g*-mode, *see* mode, gravity
- G2237+305, 332
- Gaia hypothesis, 67
- GAIA mission, 47, 252
- galactic nucleus
- active, *see* active galactic nucleus
- galaxies
- interacting, 47
- galaxy, 18, 46, 47, 90, 98, 103, 104, 120, 121, 131, 135, 149, 162–164, 164, 165–180, 212, 260, 263, 268, 270, 279, 289, 292, 293, 299–301, 305, 309, 321, 322, 324, 325, 330–343, 345, 347–357, 359–361, 363–368, 370, 371, 373, 375, 376–381, 383, 384, 390, 391, 393, 394, 397, 398, 401, 402, 404–407, 412, 417, 424–434, 436, 438, 440
- active, *see* active galaxy
- center, 163, 164, 166, 170, 285, 289, 302, 309, 321, 322, 339, 345, 346
- classification, 166
- cluster, 107, 120, 164, 170–172, 314, 317, 331, 332, 342, 343, 347, 366, 367, 373, 375, 376–378, 404, 406, 428–430, 439
- collisions, 166
- density wave in, 166, 426
- elliptical, *see* elliptical galaxy
- formation, 131, 173, 175, 177, 331, 368, 377, 378, 381, 382, 394, 402, 405, 406, 413, 424, 425, 429, 430, 432
- seeds, 133, 159, 173, 391, 405, 406, 430
- standard model, 424
- irregular, *see* irregular galaxy
- mass, 168–170, 173, 340
- mass estimate, 169
- measuring distance to, 168
- recession speed, 167, 360, 363
- satellite, 169
- spectrum, 167
- spiral, *see* spiral galaxy
- Galaxy, ours (the Milky Way), 47, 48, 103, 104, 124, 128, 131, 132, 134–136, 149, 160, 163, 164, 164, 165, 167–170, 172, 174, 175, 188, 189, 201, 207–209, 221, 261, 268, 270, 271, 273, 279, 293, 294, 299, 309, 319, 321, 326, 332, 338, 339, 342, 345, 347, 360, 363, 394, 406, 407, 412, 413, 423, 424, 428, 429, 431, 434, 436
- center, 47, 168–170
- mass, 168, 169
- size, 168
- Galileo Galilei, 2–7, 9–11, 13–15, 18, 20, 23, 25, 46, 69, 163, 179, 181–183, 184, 188, 189, 202, 211, 212, 214, 222, 417, 437
- Galileo spacecraft, 60, 61
- gallium, 130
- gamma-ray, 87, 95, 96, 133, 201, 206, 278, 279, 305, 354, 406, 429, 437, 441
- bursts, 279, 294, 305, 347, 407, 429
- luminosity, 279
 - timescale, 279
- gas
- collisionless, 164, 424
 - degenerate, *see* degenerate gas
 - interstellar, 133, 134, 166, 426
 - cloud, *see* molecular cloud
 - plasma, 271
- intracluster, 171
- general relativity, 11, 12, 14, 16, 18, 37, 38, 40, 49, 88, 146, 148, 149, 156, 171, 178–180, 184, 187, 190, 194, 210–213, 223, 225, 236–241, 243–245, 247, 250–254, 256, 258–261, 264–268, 276, 280–282, 285, 286, 288–290, 292, 295, 296, 302, 309–312, 312, 313, 315–319, 323, 331, 346, 359, 360, 363, 366, 383, 384, 387, 391, 394, 395, 400, 402, 404, 405, 407–410, 421, 424, 425, 427–430, 433, 434, 436–439, 441, *see also* Einstein, field equations of general relativity
- tests of, 153, 236, 244, 318, 319, 323
- GEO600, *see* interferometric detector, GEO600
- geodesic, 223, 224, 231, 232, 234, 235, 249, 251, 292, 312, 343, 405, 429
- geologist, 122, 123, 399
- geometry, 22, 23, 26, 36, 38, 111, 156, 198, 213, 217, 219, 222–224, 226–228, 232, 234–236, 239–241, 248, 249, 252, 256, 260, 286–288, 295–298, 302, 311, 317, 320, 322, 324, 333, 336, 337, 348, 384, 385, 387–390, 402, 405, 424, 427, 428, 432
- geostationary, 19
- giants
- standing on shoulders of, 313
- global, 40, 119, 230, 349, 404, 429, 432, *see also* local properties, 349
- Global Positioning System (GPS), 16, 17, 183, 198, 203, 204, 230, 289
- globular cluster, *see* star, cluster, globular center, 158, 162
- God, 178, 345, 392
- Goddard Space Flight Center (GSFC), 253, 332, 354, 381, 391
- GP-B, *see* Gravity Probe-B
- GPS, *see* Global Positioning System
- grand unified theory (GUT), 127, 400, 401, 423, 429, 440
- grandmother (grandfather) paradox, 415
- gravitation, *see* gravity
- gravitational collapse, 89, 98, 101, 133, 134, 137, 146, 148–151, 160, 172, 173, 176, 178, 207, 212, 239, 240, 243, 260, 262, 263, 265, 266, 268, 270, 273, 277, 279, 282, 283, 290, 293, 294, 296, 303, 307, 314, 318, 324, 350, 377, 378, 411, 412, 415, 429, 431, 440
- gravitational constant *G*, 13, 31, 43, 134
- gravitational contraction, 122, 139
- gravitational field, 11, 15–17, 23, 34, 39–41, 48, 63, 64, 74, 101, 136, 153–155, 157, 160, 162, 175,

- 180, 203, 213, 219, 222, 224, 225, 231, 232, 234, 238–245, 247–249, 251, 256–259, 261, 271, 280, 281, 285, 287, 288, 290, 292, 294, 296, 309–311, 317, 318, 320, 322, 329, 341, 342, 358, 360, 365, 379, 380, 383, 384, 391, 396, 407, 408, 426, 429, 434, 436, 437, 440
Newtonian, 149, 212, 231, 235, 243, 245, 248, 249, 280, 298, 314, 315, 332
spherical implies static, 243, 290, 311
gravitational lens, 34, 171, 172, 308, 331–343, 379, 380, 392, 405, 427, 429, 434, *see also* image; lens
diverging, 334
magnification, 172, 332, 333, 337, 339
microlensing, 335, 338, 339, 342, 343, 434
duration, 342
odd number of images, 333, 341
second direct image, 335
time-delay, 342
gravitational radiation, *see* gravitational wave
Laplace model for, 310
gravitational radius, 36, 37, 286, 295, 300, 320, *see also* black hole, horizon
gravitational slingshot, 51, 57, 63, 155, 234, 429
gravitational wave, 12, 39–41, 120, 123, 143, 162, 184, 186, 191, 212, 221, 239, 259, 260, 274, 275, 277, 279, 284, 292, 294, 296, 308–312, 312, 313–317, 319–330, 346, 355, 359, 405–407, 409, 410, 417, 422–425, 429–431, 435–437
how radiated, 313
weakness of, 317
gravitational wave astronomy, 212, 314, 330, *see also* gravitational waves (in astronomy)
gravitational wave detector, 77, 84, 143, 165, 260, 312, 319, 323–326, 329, 347, 355, 422, 431
bar, *see* bar detector
external vibration, 328
interferometer, *see* interferometric detector
gravitational waves cosmological, *see also* gravitational waves, sources, Big Bang Universe transparent to, 325
gravitational waves (in astronomy) cosmological, 330
radiation reaction, 12, 319, 437, *see also* Hulse–Taylor binary pulsar
sources, 310, 313–317, 319, 328, 329, 425
Big Bang, 325
neutron star, 314
normal modes, 325
pulsars, 324
relativistic, 243
surveys, 325
gravitational waves (theory) action on free particles, *see* gravitational waves (theory), polarization
action on matter, 310, 311, 312
amplitude, 312, 312, 314, 319
and strong gravity, 317
early doubts about, 313
energy conservation, 317
definition, 317
energy flux, 316
frequency, 319–321, 325
how determined, 314
in other theories, 309
luminosity, 319, 320
polarization, 311
ellipses, 313, 316
transverse, 311
scalar waves, 312
spectra, 314
gravitoelectric aspect of gravity, 243, 245–248, 250, 254, 256, 429
gravitomagnetic aspect of gravity, 239, 240, 243, 245–253, 255, 256, 297, 298, 426, 429
in spacetime-interval, 251
inverse cube law, 251
graviton, 144, 191, 308, 410, 430
gravity, 31, 134, 164, 212, 217, 219, 224, 256, 260, 308, 407, 417, 419, 431, *see also* gravitational field
acceleration of, 3, 6, 10, 16, 19, 21, 22, 27, 31, 34, 35, 39, 41, 53, 73, 79, 91, 92, 94, 156, 157, 169, 222, 224, 231, 243, 245–248, 250, 291, 310, 339, 358, 360, 363, 419, 421, 429, 440, 441
affects time, 18, 415, *see also* curvature, time
engine of the Universe, 309
hidden, 404
law of, 11–13, 27, 28, 33, 173, 363, 404
modifications, 49, 404
Newtonian, 14, 18, 29, 31, 163, 168, 232, 259, 358, 400, 409, 431
Newtonian, 14, 38, 48, 49, 57, 147, 178–180, 211, 212, 224, 225, 229, 230, 232, 234, 236, 237, 239–241, 243, 244, 246, 247, 257, 259, 261, 265, 266, 280–282, 290–292, 309–311, 320, 321, 323, 359, 360, 366, 383, 384, 391, 421, 429, 431
no shield for, 309
noise, 329
of a spherical mass, 154
relativistic, 148, 179, 180, 193, 194, 211, 212, 225, 236, 242, 247, 253, 256, 259, 261, 309
law, *see* general relativity
sources, 240
repulsive, 358
scalar law of gravitation, 312
scalar–tensor theory, 312
short-range law, 409
time-independent, 63
weakness of, 309, 317
Gravity Probe-B (GP-B), 252, 253
great circle, 35, 213, 222–224, 385, 390, 429
Greek
astronomy
distance to Moon, 104
mythology, 25, 26, 104
greenhouse effect, 66–69, 113, 430
greenhouse gases, 67, 114, 430
GRO J0422+32, 299
GRO J1655-40, 299
GRS 1124-683, 299
GS 2000+250, 299
GS 2023+338, 299
GSFC, *see* Goddard Space Flight Center
GUT, *see* grand unified theory
Guth, A, 401
gyroscope, 251, 252
H1705-250, 299
half-life, 126, 197, 198, 201, 430
Hanford, Washington, 327
Hannover, Germany, 327

- hard-core, *see* force, nuclear, hard-core repulsion
- Hawking radiation, 298, 304–308, 322, 408, 422, 430, 437
temperature, 305, 307, 308, 430
- Hawking, S W, 290, 304, 305, 307, 308
- HDF-S team, 353
- Heisenberg uncertainty principle, 83, 86, 142, 143, 260, 289, 295, 328, 391, 395, 396, 407, *see also* quantum uncertainty
- Heisenberg, W, 83, 84, 142, 143, 289, 295, 395
- helioseismology, 98, 99, 102, 129, 284, 430
- helium, 65, 69–73, 78, 79, 97, 121–127, 131–133, 138, 140, 143, 147, 151, 159, 172, 177, 206, 210, 212, 322, 327, 362, 367, 370, 372–374, 375, 376, 381, 393, 398, 411
nucleus, 123–125, 127, 132, 133, 370, 372, 435
- hertz, 87, 316
- Hertzsprung, E, 139
- Hertzsprung–Russell diagram, 139, 164, 430, 433
- Hewish, A, 271, 272
- hidden mass, *see* dark matter
- Hipparcos mission, 105, 140, 153, 252
- Hitler, A, 84, 112, 286
- Homestake Gold Mine, 128
- homogeneity, 32, 71, 345, 347–350, 355, 357–361, 363, 365, 367, 368, 372, 377, 378, 383–387, 389, 391, 393, 394, 397, 428, 430–432
- homogeneity and isotropy problem, 384, 401, 430
- Hooke, R, 13
- horizon, *see* black hole, horizon; particle horizon
- hot dark matter, 382, 430, *see also* cold dark matter (CDM)
- Hoyle, F, 124, 125, 132, 134, 358, 422, 423, 439
- HST, *see* Hubble Space Telescope
- Hubble constant, 167, 333, 342, 348, 350–352, 357, 358, 360, 361, 363, 364, 366, 371, 373, 374, 388–390, 401, 413, 419, 425, 430, 431
- expansion law, 167, 168, 175, 345, 348, 349, 355, 360–364, 388
- expansion speed, *see* Hubble, constant
- time, 352, 364, 419, 431, *see also* Universe, age of
- Hubble Deep Field, 353
- Hubble expansion, 349
- Hubble Space Telescope (HST), 19, 47, 164, 166–168, 176, 177, 228, 331, 332, 353, 393
- Hubble, E, 167, 168, 175, 256, 348, 349, 360, 363, 364, 374, 394, 395, 402
- Hulse, R, 317, 319
- Hulse–Taylor binary pulsar PSR1913+16, 169, 272, 318–321, 404
- hydrogen, 65, 68–70, 75, 78, 89, 97, 116, 119–127, 130, 131, 133, 135, 137–140, 143, 147, 151, 159, 170–172, 175, 177, 206, 210, 212, 322, 354, 355, 362, 363, 367, 373, 375, 426, 429, 433
ionization, 109, 375, 376
- hydrostatic equilibrium, 72, 73, 81, 94, 95, 280, 282
- hyperbola, 221, 389, 431
invariant, 221, 431
- hyperboloid, 386, 388, 389, 431
- Hyperion, 91
- hypernova, 279, 429, 431
- hypotenuse, 3, 5, 22, 431, 437
- ideal gas, 143, 146
law, 77, 78, 81, 92, 95, 280
- image, *see also* gravitational lens; lens
- direct, 335, 339–341, 426
inverted, 341
location, 333, 340
mirror, 246
- impact parameter, 37, 38, 338
- Inco Mining Company, 130
- independent motions, 4, 6, 23
- index, 92–95, 101, 110, 137, 144, 146, 156, 258, 424, 431, 440
- inertia, 10, 190, 191, 193, 194, 254, 255, 281, 431
of pressure, *see* pressure, inertia of origin of, 253
- inertial mass, 14, 190, 193, 204–206, 254, 255
density, 193, 194, 255, 257, 280
of light, 14
- inflation, 193, 239, 244, 253, 260, 353, 362, 371, 378, 381, 384, 390, 392–394, 397–402, 404–407, 409, 413, 414, 421, 426, 430, 431, 433, 434, *see also* cosmology
- infrared, 67, 87, 88, 109, 111, 113, 114, 136, 161, 172, 328, 430, 431, 434
- inhomogeneous, 39, 389, 393, 394, 414, 431
- innermost stable circular orbit, *see* orbital motion, relativistic, innermost stable orbit
- instability, 69, 101, 126, 129, 137, 138, 143, 146, 188, 197–199, 246, 254, 263, 264, 282, 283, 372, 402, 422, 431, 432
- interaction, *see* force
- interference, 186, 187, 328, 431
- interferometer, 184–187, 199, 327–330, 431, *see also* interferometric detector
- radio, 332
- interferometric detector, *see also* gravitational wave detector
- AIGO, 328
- GEO600, 327, 328
- LIGO, 322, 327, 328
- LISA
- proof mass, 330
- LISA, 309, 319, 322, 324, 329, 330
- NAUTILUS, 327
- present projects, 327
- TAMA, 327, 328
- VIRGO, 327, 328
- interstellar cloud, *see* molecular cloud
- interstellar medium, *see* gas, interstellar
- interstellar plasma, *see* gas, interstellar, plasma
- intrinsic brightness, *see* brightness, intrinsic
- invariance, 64, 182, 184, 185, 189, 195, 202, 213, 217, 220, 241, 254, 255, 257, 410, 413, 421, 424, 431, 434, 438
- invariant hyperbola, *see* hyperbola, invariant
- inverse-square law, 13, 14, 27, 49, 431
- Io, 39, 46
- ion, 67, 301, 431
- ionization, 123

- ionized, 81, 82, 97, 168, 302, 325, 354, 355, 375–378, 431
 gas, 81, 119, 163, *see also* plasma
 iron, 2, 13, 65, 125, 127, 143, 148, 151, 154, 172, 210, 267, 350
 irregular galaxy, 166, 431
 isothermal, 79–82, 431
 isotope, 125–128, 284, 393, 426, 431
 isotropy, 71, 72, 89, 242, 244, 245, 254, 357, 358, 383–387, 389, 391, 397, 421, 432
 iteration, 30
 Jeans length, 136, 137, 432
 Jeans, J, 136–138, 432
 jet, 74, 163, 164, 174–177, 210, 250, 251, 267, 279, 299, 349, 432
 Jet Propulsion Laboratory (JPL), 26, 39, 60, 61, 68, 309
 Jodrell Bank Radio Observatory, 331
 joule (unit for energy), 54, 88, 315, 432
JPL, *see* Jet Propulsion Laboratory
 Jupiter, 7, 25, 28, 29, 32, 33, 38, 41, 45, 46, 48, 51, 52, 56–63, 65, 66, 90, 155–159
 mass of, 59, 338
 moons, 7, 25, 46
 orbit, 7, 33, 48, 56, 59, 60
 tidal force, 48
 Kamiokande, 129, 130
 SuperKamiokande, 130
 kelvin, 75, 116, 117, 119, 171, 426, 432
 Kepler's
 constant, 29, 291
 first law, 32, 56, 58, 291
 second law, 32
 third law, 27, 29, 32, 33
 Kepler, J, 13, 25, 27–29, 32, 33, 56, 58, 291
 Kerr geometry, 297, 298, 302, 322
 Kerr, R, 219, 297, 298, 302, 307, 439
 kick, 51, 160, 269, 270
 kiloparsec (kpc), 104, 432
 Kuiper Belt, 61, 66, 432, 435
 Kuiper, G, 61
 LAGEOS satellites, 252, 253
 Landau, L, 268, 269, 316
 Laplace, P, 36, 37, 287, 310, 319
 Large Magellanic Cloud (LMC), 129, 149, 338, 339
 laser, 135, 145, 187, 328, 433
 latitude, 26, 35, 36, 83, 213, 226, 228, 424, 432
 launch speed, 24, 52, 56, 57
 launch window, 57
 law of sines, 26, 432
 Le Verrier, U, 48, 49
 lead, 210, 264
 Leaning Tower of Pisa, 14
 Leibniz, G, 9, 423
 length, 104, 113, 146, 181, 186–188, 192, 194, 199, 200, 200, 201–203, 207, 219, 220, 229, 286, 326, 327, 348, 349, 385, 397, 426, 432, 433, 437–440
 contraction, 185, 189, 195, 201, 428, *see also* Lorentz–Fitzgerald contraction
 measuring, 186, 199, 202, 438
 proper, *see* proper distance; proper length
 lens, *see also* gravitational lens; image
 converging, 339
 diverging, 333, 335, 339, 343, 426
 Lense, J, 252
 Lense-Thirring effect, 251–253, 287, 297, 426, 430
 lepton, 127, 128, 130, 406, 422, 427, 432, 434, 441
 families, 127, 434
 number, 128, 130, 296, 432
 conservation of, 130
 life, 2, 54, 62, 65, 67–69, 75, 77, 100, 101, 103, 112, 120, 124, 127, 131–135, 138–141, 146, 151, 183, 190, 210, 263, 266, 289, 305, 345, 349, 367, 370, 392, 394, 408, 410–413, 417, 435
 evolution of, 121, 122, 131, 132, 145, 345, 410–414
 Lifshitz, Y, 316
 light, 2, 14–18, 20, 34, 36–38, 54, 66, 68–70, 82, 85–89, 95, 96, 105–107, 109–119, 121, 123, 124, 127, 130, 133–137, 142, 145, 149, 150, 153, 154, 163, 168–172, 174–176, 178–189, 191, 192, 196–198, 201, 202, 206, 207, 210, 212, 215, 217, 219, 220, 225, 230, 232, 235–239, 245, 252, 253, 256, 259–261, 266, 270–272, 277–279, 285, 287, 288, 293–295, 297, 301, 304, 309, 311, 314–317, 324, 327–330, 332–343, 347, 348, 350, 353, 355, 356, 359, 364, 365, 372–374, 379, 390, 395, 396, 404–406, 421–424, 426, 428–432, 434–440, *see also* electromagnetic waves
 deflection, *see* deflection of light
 energy of, 85
 frequency, 15–17, 191, 266, 436, *see also* color, of light
 speed, 12, 14, 36, 54, 86, 87, 96, 97, 111, 113, 123, 124, 131, 134, 145–148, 153, 163, 164, 174, 175, 179–192, 195–199, 201–203, 205–211, 220, 228, 233, 246, 260, 264–267, 272, 275, 288, 294, 295, 297, 298, 308–310, 318, 323, 346, 359, 360, 364, 365, 370, 382, 383, 389, 397, 406, 408, 410, 419, 424, 429, 430, 440
 as limit, 187, 191
 invariance, 187, 197
 light deflection, *see* deflection of light
 light-cone, 220, 285, 289, 297, 336, 347, 432, *see also* spacetime structure
 future, 220
 past, 220, 346, 347, 383, 389, 393
 light-meter, 219, 221, 243
 lightlike, 190, 389, 432
 LIGO, *see* interferometric detector, LIGO
 limestone, 67
 linear momentum, *see* momentum
 linear relationship, 97, 107, 240, 241, 432, 435
 LISA, *see* interferometric detector, LISA
 lithium, 127, 372, 373
 little green men, 25, 271
 Livingston, Louisiana, 327
 LMC, *see* Large Magellanic Cloud
 LMC X-3, 299
 local, 38, 40, 44, 94, 103, 129, 149, 222–224, 226, 227, 230–234, 251, 255, 257, 260, 280, 288, 289, 293, 295, 297–299, 303, 313, 317, 347, 349, 358, 360, 366, 383, 389, 400, 404, 428, 429, 432, 437, 439, *see also* global
 flatness, 223, 224, 229, 231, 232, 258, 289, 424, 429, 432, 436
 and free-fall, 224
 properties, 255, 349
 Local Group (of galaxies), 170, 347, 363, 429

- logarithm, 13, 107, 108, 141, 306, 308, 424, 428, 432
 natural, 119
 logarithmic scale, 106–108, 110, 432
 London, 213
 longitude, 35, 36, 213, 226, 228, 251, 252, 433
 longitudinal, 311, 422, 433, 436
 loop, 5, 81, 85, 251, 252, 349, 405, 433, 439
 Lorentz, H A, 185, 186, 188, 199
 Lorentz–Fitzgerald
 contraction, 185, 188, 192–194, 199, 201–203, 209, 218, 235, 247, 250, 257, 281, 422, 433,
 see also length, contraction
 transformation, 190, 213, 433
 luminosity, 105–108, 116–118, 122, 123, 128, 129, 133, 138–141, 148, 151, 161, 168, 169, 172, 174, 176, 265, 267, 273, 275, 277, 279, 294, 300, 305, 319–321, 324, 350, 419, 421, 427, 430, 433, 438
- M**₁, 150
M₃, 165
M₁₆, 136
M₃₁, 104, 168, 363, 434
M₃₂, 168, 170
M₃₃, 170
M₈₇, 164, 166, 170, 174, 176
 Mach number, 254
 Mach, E, 251, 253, 254
 Machian, 253
 MACHO, *see* massive compact halo object
 macroscopic, 89, 296, 303, 307, 308, 313, 328, 401, 433, 439
 Magellanic Clouds, 47, 168, 170, 299, 428
 magnetar, *see* pulsar, magnetar
 magnetic field, 81, 82, 90, 98, 99, 128, 138, 142, 177, 181, 245, 248, 261, 268, 272–274, 283, 284, 299, 375, 399, 408, 427, 433, 437, 439, 441
 interstellar, 406
 pulsar, *see* pulsar, magnetic field
 magnetic force, 12, 212, 263, 433, 440
 magnetic monopole, 398, 401, 407, 433
 magnetism, 180, 182, 240, 243, 245, 248, 398–400, 427, 429, 433
- magnitude, 11, 16, 23, 29, 106–110, 140, 180, 248–250, 291, 315, 421, 422, 424, 433, 440
 absolute, 106, 108, 116, 352, 421,
 see also brightness, intrinsic
 apparent, 106, 108, 315, 316, 352, 422, 433, *see also* brightness, apparent
 blue, 110, 424
 bolometric, 109, 422
 visual, 109, 110, 140, 424, 440
 main sequence, 118, 139, 140, 153, 430, 433
 Manhattan, 266
 map
 Earth, 223
 MAP satellite, 381, 391
 Mars, 13, 25, 26, 28, 46, 51, 56, 58, 66, 68, 69, 82, 83, 90, 159, 228, 422, 440
 life on, 69
 moons, 90, 91
 water on, 26, 69
 maser, 135, 433
 mass, 9, 10, 13–15, 22, 27–29, 31–38, 43, 48, 52–54, 59, 62, 65, 69, 70, 72, 73, 76, 78–82, 84, 88–95, 97, 100, 101, 103, 105, 107, 114, 118, 122–124, 128, 131, 133, 134, 136–149, 151, 153–157, 159–163, 164, 165, 166, 168–176, 179, 180, 184, 187, 188, 190–195, 200, 204–207, 210, 222, 231, 240–255, 257, 260–266, 269, 275–283, 286, 289–297, 299, 300, 304, 305, 307, 308, 314, 315, 318–323, 330, 331, 333, 336, 337, 339, 341–343, 350, 351, 357–364, 373, 374, 377, 378, 380–383, 388, 398–400, 403–405, 408–412, 415, 419, 421–423, 426, 427, 429–436, 438, 440, 441,
 see also inertial mass
 active curvature, *see* active curvature mass
 active gravitational, *see* active gravitational mass
 atomic, *see* atomic mass
 density, 144, 180, 193, 194, 244, 245, 247, 250, 255, 257, 280, 281, 356, 360–362, 364, 373, 374, 380, 388–390, 392, 397, 398, 400, 403, 413, 425, 439, *see also* energy, density; mass-energy, density
 dependence on speed, 187, 190, 205, 250
 in astronomy, 276
 mass function, 156, 433
 mass-energy, 145, 179, 204, 241, 246, 247, 257, 281, 315, 316, 323, 362, 374, 376, 400, 421
 conversion, 122
 density, 247, 257, 425, *see also* energy, density; mass, density equivalence, 190, 195, 204
 mass-to-light ratio, 169, 433
 massive compact halo object (МАЧО), 172, 173, 335, 338, 339
 massless particle, 131, 308
 matrix, 258, 424, 434, 440
 matter–antimatter asymmetry, 369
 Max Planck Institute for Gravitational Physics (Albert Einstein Institute, AEI), 239, 285, 292, 323
 Maxima, *see* cosmic microwave background (CMB, Maxima experiment)
 Maxwell’s equations, *see* electromagnetism, Maxwell’s equations
 Maxwell, J C, 76, 182, 185, 427, 440
 MCG-6-30-15, 300–302
 MDI instrument, 98, 102
 mean free path, 97
 mean molecular weight, 81, 92, 94
 measurable, 17, 28, 31, 48, 53, 83, 113, 154–156, 179, 184, 185, 230, 233, 237, 273, 280, 286, 289, 296, 297, 321, 349, 353, 359, 364, 379, 389, 429, 433
 mechanics, 6, 9, 11, 14, 26, 76, 84, 112, 180, 182, 183, 191, 193, 206, 219, 224, 282, 285, 306, 308, 432, 434, 439
 megaparsec (Mpc), 104, 163, 167, 419, 434
 Mercury, 7, 14, 28, 29, 32, 37, 46, 48, 49, 51, 55, 61, 66, 83, 155–158, 160, 211, 225, 236–238, 247, 318, 440
 perihelion, 48, 211, 225
 precession, 48, 49, 225, 237, 238, 285
 rotation, 46
 mesosphere, 82
 Messier, C, 168
 methane, 67, 430
 metric tensor, 257, 258, 434, 438
 Michell, J, 36, 37, 287
 Michelson, A A, 184–187, 199

- Michelson–Morley experiment, 184, 186–188, 199
- Mickey Mouse coordinate system, 213
- microlensing, *see* gravitational lens, microlensing
- micron (μm), 87, 117, 426, 434
- microscopic, 88, 196, 242, 307, 308, 433
- microwave, 137, 163, 177, 325, 353–358, 360, 362, 364, 365, 369, 370, 374, 376, 378–382, 384, 390–393, 402, 403, 406, 425, 434, 439
- Milky Way, *see* Galaxy
- Minkowski geometry, *see* spacetime geometry, flat
- Minkowski spacetime, *see* spacetime, Minkowski
- Minkowski, H, 219
- mirror, 186, 187, 196–198, 200, 200, 201, 202, 249, 253, 327–330, 332, 356, 396, 426, 440
- missing mass, *see* dark matter
- mode, 100, 101, 425
- frequency, *see also* frequency, characteristic
 - fundamental, 99, 101, 271
 - frequency, 99, 100, 428, 435
 - gravity, 102
 - normal, 100, 435
 - overtone, 99, 100, 435
 - pressure, 101
 - vibration pattern, 99–102, 435
- molecular cloud, 65, 121, 131, 133, 266, 431
- molecular clouds, 135–138, 166, 412, 433, 434
- molecular weight, 81, 92, 94
- molecule, 67, 69–71, 76, 78–81, 83–85, 87, 88, 90, 91, 114, 122, 134, 135, 137, 164, 194, 242, 257, 306, 311, 365, 378, 399, 423, 433–435, 439
- momentum, 52–54, 64, 83, 84, 113, 142–146, 157, 190, 191, 204–207, 240–242, 248–250, 256, 258, 295, 296, 313, 424, 427–429, 434
- angular, *see* angular momentum
- monster (at the center of a quasar), 175–177
- Moon, 10, 12, 14, 25, 26, 26, 27–29, 31, 33, 38–46, 48, 49, 51, 65, 66, 103, 104, 126, 153, 207, 223,
- 252, 310–312, 315, 317, 422, 426, 440
- center, 45
- distance, 26
- measured by Greeks, *see* Greek astronomy, distance to Moon
- mass of, 27, 91
- orbit
- period, 42
 - plane, 43
 - radius, 27
- radius, 45
- solid tides, 44
- tidal effect on Earth, *see* tide, due to the Moon
- Morley, E, 184, 186
- motion
- in independent directions, *see* independent motions
 - laws of, *see* Newton, laws of motion
- multiverse, 392
- muon, 199, 209, 374, 406
- neutrino, *see* neutrino, mu-type
- Narlikar, J, 125, 358
- NASA, *see* National Aeronautics and Space Administration
- National Aeronautics and Space Administration (NASA), 19, 26, 39, 47, 60, 61, 68, 85, 90, 98, 102, 136, 142, 149, 153, 163, 164, 165–167, 177, 225, 228, 252, 267, 278, 279, 300, 309, 329, 331, 332, 353, 354, 381, 391
- National Geographic Society's Palomar Observatory Sky Survey, 104
- National Science Foundation (NSF), 150, 176, 328
- natural selection, 134
- NAUTILUS, *see* interferometric detector, NAUTILUS
- NCSA, 239, 292
- nebula, 103, 136, 140, 142, 148, 150, 159, 167, 168, 267, 270, 271, 434, 436
- Neptune, 25, 28, 29, 32, 49, 51, 61, 66, 432
- neutrino, 89, 93, 102, 123, 127–131, 133, 147, 149, 150, 172, 173, 188, 191, 205, 206, 370, 372–375, 375, 376, 382, 392, 399, 411, 412, 422, 427, 430, 432, 434, 435, 438, 441
- astronomy, 123
 - electron-type, 127, 130, 374, 427
 - mass, 412
- mass, 173
- mu-type, 374
- oscillation, 130, 131, 173
- tau-type, 374
- three types, 374, 434
- neutrinos
- from SN1987A, 150
 - solar, *see* solar neutrinos
- neutron, 81, 91, 121, 122, 125–127, 132, 143, 144, 146–149, 172, 177, 193, 210, 260, 262–264, 268, 275, 280–282, 284, 368, 369, 372–374, 375, 411, 412, 419, 422, 423, 425, 426, 428, 431, 434, 435, 439, 441
- decay, 172, *see also* baryon, number conservation of
- mass difference with proton, *see* proton, mass difference with neutron
- neutron star, 84, 103, 118, 120, 134, 135, 140, 147–151, 153, 154, 159–161, 165, 171, 175, 180, 193, 243, 252, 260–267, 267, 268–279, 281–284, 286, 290, 299, 300, 314, 317, 318, 320, 322, 324, 325, 339, 342, 350, 404, 410–412, 423, 425, 426, 429, 433–435, 437, 439–441
- abundance, 261
- binary, 165, 265, 282, *see also* Hulse–Taylor binary pulsar
- coalescence, 279
- binary pulsar, 169, 272, 275, 276, 318–320
- cluster, 175
- crust, 275, 284, 324, 425, 429
- magnetic field, 268
- mass, 148, 160, 264–266, 276, 277, 283
- maximum mass, 148, 160, 175, 264, 266, 282, 293
- minimum mass, 264
- seismology, 284
- starquake, 275
- structure, 212, 275, 281, 284, 325
- vibrations, 284
- New General Catalog (NGC), 168
- New York, 103, 135, 213, 226, 428
- Newton
- constant, *see* gravitational constant G
 - law of gravitation, *see* gravity, law of, Newtonian

- laws of motion, 6, 9, 12, 18, 36, 53, 179, 431, 432
 first, 9, 10, 23, 32, 56, 58, 223, 291, 432
 second, 9–11, 29, 32, 53, 54, 73, 190, 204, 427, 432, 438
 third, 11–14, 27, 29, 32, 33, 54, 78, 162
 third law loophole, 12
 newton (unit for force), 10, 432
 Newton, I, 6, 9–15, 25–29, 31, 33–36, 44, 53, 73, 94, 145, 163, 173, 177–181, 183, 184, 189, 201, 202, 211, 214–216, 224, 225, 238, 240, 241, 254, 260, 261, 310, 359, 414, 417, 423, 432
NGC, *see* New General Catalog
NGC3242, 142
NGC4038, 167
NGC4039, 167
NGC4414, 164, 166
 Niagara Falls, 398
 nitrogen, 65, 66, 69, 71, 78, 83, 121, 198, 210, 302
NOAO, 150, 176, 270
 Nobel Prize
 Physics, 89, 123, 131, 145, 272, 319, 355, 395
 non-linear, 240, 258, 292, 316, 426, 427, 435
nova, 103, 125, 153, 159, 160, 299, 300, 435
NSF, *see* National Science Foundation
NSSDC, 47
 nuclear bomb, 149, 151, 210, 269, 350
 hydrogen bomb, 85, 137, 138
 nuclear energy, 93, 125, 127, 135, 151, 174, 417
 nuclear force, *see* force, nuclear
 nuclear fusion, 138
 nuclear physics, 121, 124, 125, 127, 129, 135, 145, 149, 153, 261, 262, 264–266, 280, 281, 283, 284, 293, 357
 nuclear reaction, 89, 94–96, 99, 102, 120–130, 132, 133, 137, 138, 141, 147–149, 151, 160, 172, 174, 241, 278, 281, 293, 296, 338, 350, 354, 372–375, 375, 382, 422, 423, 427, 434, 435, 437, 438
 chain reaction, 159, 439
 nuclear reactor, 103, 149, 190, 206
 nucleon, 262–265, 373, 374, 376, 377, 411, 435, 437, *see also* proton; neutron
 mass, 262, 373
nucleus, 70, 87, 89, 91, 97, 103, 122–129, 132, 133, 143, 146–148, 151, 176, 197, 198, 205, 206, 261–265, 275, 284, 296, 301, 350, 354, 369–372, 375, 376, 407, 410–412, 422, 423, 425, 426, 428, 429, 431, 434, 435, 437, 439
 density of, 261–263, 265, 293
 numerical relativity, 292
 observers, *see* relativity, experimenters or observers
 Occam's razor, 259, 348, 435
 Occam, W, 259, 348, 435
 ocean, 39, 41, 44, 65, 67–70, 82, 123, 207, 317, 399, 433
 temperature, 82
 tide, 39, 44
 octave, 100
 Olbers' Paradox, 177, 178, 435
 Olbers, W, 177, 178, 435
 Oort Cloud, 61, 435
 Oort, J, 61, 169
 opacity, 97
 Oppenheimer, J R, 268, 269
 orbital motion, 7, 12–14, 16, 17, 19, 20, 22–37, 40, 42, 43, 45–49, 51–62, 64, 66, 79, 81, 85, 98, 100, 105, 116, 145, 149, 153–160, 162, 163, 164, 165, 166, 168–170, 176, 177, 180, 185, 198, 203, 211, 212, 222, 231, 233, 234, 236–239, 252, 253, 257, 261, 266, 269, 270, 276–279, 286, 290–292, 297–299, 302, 309, 310, 317–322, 329, 339, 358, 364, 405, 410, 417, 421–424, 426, 431–433, 435, 439, 440
 around Sun, 23, 27, 48, 52, 54, 57, 58, 104, 155, 185, 222, 237, 309, 329, 358, 419, 422, 435
 binary stars, *see* binary stars, orbit
 circular, 19, 21–23, 25–27, 53, 62, 154, 156–159, 169, 171, 233, 237, 265, 269, 291, 297, 319–321, 323, 422, 431
 speed, 23, 35, 53, 269, 339
 decay, 277, 279, 310, 319, 423
 elliptical, 13, 22, 26, 28, 48, 58, 61, 157, 158, 276, 277, 291, 318, 329, 358, 422, 435
 energy, *see* energy, orbital
 hyperbolic, 358
 parabolic, 358
 plane of orbit, 43, 66, 154, 277
 radius, 22, 56, 59, 156, 170, 262, 319–321
 relativistic
 innermost stable orbit, 291, 431
 order-of-magnitude, 88, 91
 organ, 21, 99, 100
 Orion nebula, 103
 outgassing, 66, 68, 69
 overtone, *see* mode, overtone
 oxygen, 65, 68, 69, 71, 78, 79, 91, 121, 124, 125, 127, 132, 133, 139, 147, 148, 151, 198, 205, 210, 266, 301, 302, 370, 435
 ozone, 80, 82, 435
 p-mode, *see* mode, pressure
 p–p chain, 127, 129
 panspermia, 124, 134, 435
 paradox, 71, 177, 178, 187, 198, 207, 209, 334, 359, 415, 435
 parallax, 26, 26, 104, 105, 140, 333, 435
 parsec, 104, 107, 419, 432, 434, 435
 particle
 elementary, 87, 143, 144, 172, 173, 188, 197, 199, 205, 260, 284, 313, 345, 381, 391, 392, 406, 421, 423, 428, 436, *see also* under specific particle names
 neutral, 97, 172, 206, 375, 377, 381, 434
 particle horizon, 346, 347, 359, 389, 430, 435
 pascal (unit for pressure), 72
 Pascal, B, 72
 pattern matching, 322, 325, 435
 Pauli exclusion principle, 142–147, 264, 428
 Pauli, W, 142, 205
 Penrose process, 298, 299, 435
 Penrose, R, 290, 298, 425
 Penzias, A, 355
 perfect absorber, 356, 375
 periastron, 318, 435
 perigee, 45, 435
 perihelion, 28, 29, 32, 45, 49, 51, 56, 58, 225, 237, 238, 247, 285, 318, 435
 perihelion advance

- accumulated, 237
 periodic table, 264
 permanent magnet, 399
 perpetual motion, 111, 299, 435
 perturbation theory, 48
 Phobos, 90, 91
 photoelectric effect, 86–89, 112, 436
 photon, 86–89, 93, 95–97, 102, 111,
 114–120, 128, 133, 137, 143–
 146, 151, 181, 183, 184, 185–
 188, 190, 191, 195–198, 200,
 200, 201, 202, 204, 206, 207,
 210, 219, 220, 224, 230–233,
 235–237, 241, 244, 245, 279,
 285, 287–289, 297, 298, 300–
 304, 307, 309, 328, 346, 354–
 357, 364, 365, 368–372, 374–
 380, 382, 383, 395–397, 399,
 402, 406, 408, 410, 415, 417,
 421, 423–426, 428–430, 432,
 436, 438, 440
 collision with particle, 406
 gas, 87, 89, 94, 96, 133, 354, 356,
 361, 364, 370–372, 374, 375,
 376, 377
 impossibility of decay, 188
 momentum, 145, 191, 206
 scattering, 115
 zero rest-mass, 187, 191
 photosphere, 88, 96, 115, 436
 physical law, 3, 4, 9, 14, 174, 350
 piano, 222
 Pisa, 2, 14, 328
 Planck
 density, 397, 400
 energy, 391, 400
 length, 146, 294, 295, 308, 397,
 401, 408, 409, 416, 436, 438
 mass, 143, 146, 294, 295, 305, 400,
 408, 410, 411, 436
 time, 294, 295, 375, 408, 416, 436
 Planck satellite, *see* cosmic microwave background (CMB),
 Planck satellite
 Planck's constant, 83, 86, 87, 111,
 112, 134, 142, 143, 204, 260,
 280, 295, 307, 308, 315, 397,
 408, 410, 419, 423, 428, 436
 Planck, M, 83, 86, 109, 111–113,
 115–117, 119, 146, 295, 304,
 354, 381, 409, 417, 436
 plane of the sky, 156, 314, 320, 436
 planetary nebula, 142, 148, 436
 planetesimal, 436
 planets, 6, 7, 12–14, 22, 25, 27–29,
 31–33, 42, 43, 45, 48, 49, 51,
 53, 56, 57, 60, 61, 65–67, 69, 70,
 82, 85, 90, 98, 99, 131, 133–135,
 153–155, 158, 159, 163, 168,
 185, 207, 210, 211, 233, 237,
 243, 257, 261, 276, 286, 290,
 296, 331, 347, 369, 377, 412,
 413, 417, 421, 423, 425, 436,
 440
 formation, 61, 65, 132, 158, 432
 on other stars, 131, 155
 orbits, 25, 28, 29, 34, 61, 257, 310,
 319, 417
 cycles, 25
 epicycles, 25, 121, 425, 427
 protoplanet, 159
 terrestrial, 66, 69, 440
 plasma, 88, 89, 96, 210, 271, 273,
 309, 325, 354, 355, 375, 376,
 426, 436
 interstellar, *see* gas, interstellar
 plate tectonics, 69, 131, 399, 436
 Pluto, 28, 29, 32, 61, 66, 103, 257,
 435
 point mass, 34, 35, 341, 361, 436
 polarization, 308, 311, 312, 316, 320,
 322, 328, 436
 Polaroid, 311
 polytrope, 92–95, 101, 144, 146, 282,
 436
 polytropic exponent, *see* polytropic index
 polytropic index, 92, 95, 101, 137,
 144, 146, 283
 positron, 124, 127, 129, 133, 368,
 369, 372, 375, 422, 436
 post-Newtonian approximation, 238,
 436
 Potsdam, 286
 Pound, R V, 16
 Pound–Rebka–Snider experiment,
 16, 17
 power, 12, 39, 43, 54, 92, 93, 97, 99,
 101, 111, 116, 117, 119, 122–
 125, 127, 137, 141, 146, 161,
 175, 182, 206, 210, 239, 271,
 274, 278, 299, 315, 316, 320,
 343, 356, 371, 394, 400, 419,
 425, 433, 434, 436, 439–441
 power law, 92, 93, 144, 282, 433, 436
 precession, 48, 49, 236–238, 252, 253,
 261, 276, 291, 318
 pressure, 44, 65, 69, 71–81, 87, 89–
 95, 98, 100, 101, 120, 133, 136–
 138, 140–148, 151, 180, 193,
 194, 206, 212, 242–244, 247,
 248, 250, 254–258, 263–265,
 280, 282, 296, 306, 309, 316,
 318, 358, 360–363, 371, 383,
 388, 395, 400–402, 405, 421,
 424, 426, 429, 432, 434–439,
 441
 atmospheric, 44, 72, 75, 77, 78, 419
 degeneracy, *see* degenerate gas
 force, 71–73, 75, 78, 79, 177, 255,
 358
 inertia of, 193
 isotropic, 72, 244, 245, 280, 388,
 405
 negative, 80, 244, 253–255, 257,
 351, 358, 378, 392, 395, 396,
 400, 401, 405, 426, 434, 440,
 see tension
 radiation, 89, 94, 128, 141, 207,
 330
 ram, 242, 247, 250, 255, 388, 437,
 439
 principle of mediocrity, 347, 424, 437
 proper distance, 220, 221, 226, 280,
 286, 287, 312, 405, 437
 proper length, 389
 proper time, *see* time, proper
 propulsion, 11
 proton, 81, 91, 97, 114, 118–125, 127,
 128, 132–134, 143, 144, 146–
 148, 172, 173, 177, 179, 180,
 196, 205, 210, 222, 260, 262–
 264, 284, 315, 328, 346, 354,
 365, 368, 369, 372–374, 375,
 378, 402, 406, 410–412, 422,
 425–428, 431, 434, 435, 437,
 439, 441
 decay, 129, *see also* baryon, number, conservation of
 half-life, 127
 mass, 131, 134, 146, 147, 265, 348,
 410–412, 419
 number in the Sun, 133
 proton–electron mass-ratio, importance of, 410
 proton–Planck mass-ratio, importance of, 410
 PSR1913+16, *see* Hulse–Taylor binary pulsar
 Ptolemy (Claudius Ptolomæus), 25, 26
 pulsar, 90, 118, 147, 150, 156, 169,
 207, 210, 261, 269–277, 279,
 283, 284, 317–321, 324, 325,
 404, 429, 432–435, 437, *see also* neutron star
 age of, 274
 by spindown, 274, 275

- dead, 273
 glitch, 275, 284, 429
 magnetar, 274, 433
 mass, *see* neutron star, mass
 millisecond, 274–276, 434
 period, 273
 spindown, 274, 324, 325
 timing observations, 276
 Pythagorean
 distance, 218, 220
 theorem, 20, 218, 226–230, 234, 427, 428, 437, 438
- QED, *see* quantum electrodynamics
 QSO, *see* quasi-stellar object
 quadratic equation, 58, 437
 quadrupole formula, 316, 437
 quantization, 113
 quantum, 83, 84, 86–88, 112, 135, 142, 238, 244, 260, 262, 264, 280, 284, 285, 289, 290, 295, 304, 307, 308, 378, 381, 391, 395, 402, 407, 408, 413–417, 423, 424, 426–428, 430, 432, 437–439
 electrodynamics (QED), 395, 396, 408, 430, 436, 437
 gravity, 120, 134, 238, 260, 290, 295, 303–305, 308, 353, 358, 391, 392, 397, 408–410, 413–417, 437, *see also* string theory; unified field theory
 measurement, 84, 416, *see also* quantum, uncertainty
 theory, 83, 84, 86–88, 97, 109, 112, 113, 116, 135, 142, 143, 145, 180, 191, 260, 261, 285, 288–290, 295, 303–305, 307, 308, 312, 315, 391, 395, 398, 407, 408, 410, 414, 415, 428, 430, 437, 441
 uncertainty, 262, 298, 304, *see also* quantum, measurement; Heisenberg uncertainty principle
 quantum fluctuation, 84, 295, 303, 304, 394, 402, 416, 437
 initial, 402
 quark, 284, 368, 370, 399, 437, 439
 matter, 284, 437
 soup, 368, 370, 375, 437
 quasar, *see* quasi-stellar object
 quasi-stellar object (QSO), 115, 161, 163, 166, 171, 175–177, 210, 212, 239, 267, 299, 331, 332, 337, 342, 343, 347, 349, 365, 375, 390, 407, 432, 437
 emission region, 176
 luminosity, 175, 176
 quintessence, 255, 402, 425, 437
 radian, 37, 38, 43, 104, 227, 237, 337, 339, 386, 437, 440
 radio galaxy, 174–176, 432
 giant, 163, 166, 174
 radio sources, 175, 268, 271
 radio lobes, 174
 radio telescope, 174, 274, 317, 321, 332
 radio waves, 12, 17, 118, 135, 163, 170, 172, 174, 268, 270–272, 318, 324, 329, 330, 437, *see also* electromagnetic waves
 transmission, 118
 radioactivity, 46, 69, 70, 95, 209, 263, 371, 422
 radioactive elements, 65, 70, 126
 radius, 19, 21, 22, 24, 26, 26, 27, 31, 33–38, 43, 56, 58, 59, 62, 72, 89–95, 97, 99, 101, 105, 107, 116–118, 140, 141, 143, 144, 146–148, 153, 156, 159–161, 165, 168–171, 174, 180, 227, 228, 237, 257, 262, 265, 266, 275, 280, 282, 283, 286–288, 290–295, 297, 300, 302, 304, 305, 312, 319–321, 335–339, 342, 349, 385–387, 389, 390, 411, 412, 419, 421, 427, 432, 435, 437, 438, 441
 circumferential, 287
 classical electron radius, 97, 419
 rain (and carbon cycle), 67, 68
 random, 65, 66, 70, 76, 76, 77, 78, 84, 85, 88, 91, 96, 97, 100, 126–128, 132, 135, 137, 143, 154, 159, 164–166, 168, 169, 180, 190, 194, 197, 242, 246, 247, 263, 265, 280, 301, 305, 306, 326, 327, 338, 339, 355–357, 366, 368, 372, 376–378, 381, 384, 393, 399, 402, 404, 405, 413, 423, 424, 430, 434, 438
 atomic motion and temperature, 122, 137, 265, 278, 301, 328
 walk, 88, 96, 97, 378, 423
 Rebka, G A, 16
 recombination, 120, 354, 426, *see also* Universe, decoupling
 redshift, 15–18, 168, 175, 191, 192, 198, 204, 206, 231–233, 243, 266, 267, 271, 277, 288, 298, 301, 318, 319, 331, 342, 351, 352, 355–357, 364, 365, 367, 368, 376, 379, 380, 390, 393, 402, 405, 437
 cosmological, 321, 355, 364, 365
 gravitational, 16–18, 38, 198, 203, 225, 230–233, 236, 241, 243, 266, 277, 287–289, 291, 301, 318, 379, 415
 Rees, M, 392
 Reines, F, 123
 relativity, 6, 14, 37, 89, 112, 120, 124, 147, 148, 153, 176, 179–181, 183, 189–193, 200, 203, 204, 206–210, 213–219, 221, 223, 224, 232–234, 238, 240, 241, 243, 244, 247, 251, 253, 261, 265, 269, 280–282, 285, 287, 290–292, 298, 301, 310, 313, 318–320, 323, 341, 357, 359, 360, 383, 386, 388, 408, 424, 427, 428, 431, 433, 440, *see also* general relativity; special relativity
 experimenters or observers, 4, 6, 7, 13, 15, 16, 37–39, 41, 105, 138, 156, 180–183, 186–190, 193, 195–204, 206–209, 213–218, 220, 224, 230–234, 239–242, 245–248, 250, 251, 253–257, 259, 266, 280, 281, 285–289, 291, 293, 296–298, 301, 303, 307, 309, 310, 313, 316, 332–334, 336, 337, 339, 347, 349, 355–357, 384, 389, 390, 395, 396, 399, 402, 410, 414, 415, 424, 426, 428–431, 433, 435–440
 general, *see* general relativity
 principle of, 6, 7, 9, 18, 23, 179, 181, 188, 193, 198, 211, 239, 254, 395, 437
 relativistic corrections, 192, 246, 280, 282
 special, *see* special relativity
 relaxation time, 164, 437
 relaxed (velocity distribution), 164, 171, 348, 438
 repulsive gravity, *see* anti-gravity
 rest frame, 301, 355–357
 preferred, 357, 399, 402
 rest-mass, 187, 190–192, 194, 196, 204–206, 210, 233, 234, 241, 247, 250, 278, 298, 299, 303, 370, 376, 427, 438

- non-zero, 196, 370
- right-hand rule, 248, *see also* two-hand rule
- rocket, 12, 21, 23, 51–54, 57, 184, 187, 188, 201, 208, 221, 241, 242, 246, 247, 250, 288, 289
- boost phase, 23
 - equation, 53, 204
- rocks
- oldest, 65, 93
- Roemer, O, 36
- roller coaster, 21
- Roman
- mythology, 12, 25
- rotation, 41, 44, 45, 70, 80, 90, 98, 99, 102, 174, 177, 251, 268, 271–273, 275, 284, 297–299, 301, 311, 317, 320, 435, 440
- axis of, 46, 98, 174, 177, 272, 320, 324
- rotation curve, 170, 438
- rubber duck, 317
- Russell, H N, 139
- sailing ship, 6, 12
- SAO, *see* South African Observatory
- satellite, 16, 17, 19, 22–24, 28, 29, 33, 45, 53, 79, 83, 98, 100, 102, 104, 105, 140, 153, 163, 168, 170, 183, 203, 252, 253, 266, 267, 271, 276–279, 289, 301, 302, 318, 351, 356, 380, 381, 391
- communications, 19, 33
 - orbit insertion, 23
- Saturn, 25, 27, 28, 33, 51, 58–61, 65, 66, 83, 161
- moons, 82, 91
- scalar, 22, 438, *see also* vector
- scale-height, 81, 438
- Schmidt, M, 175
- Schrödinger, E, 84, 395
- Schwarzschild geometry, *see* black hole
- Schwarzschild, K, 219, 285–287, 295, 305
- Schwarzschild, M, 286
- Schwinger, J, 395
- SCO X-1, 271
- selection effect, 166, 438
- Severn river, 39, 44
- Seyfert galaxy, 163, *see also* active galaxy
- nucleus, 176
- shelled animal, 67
- shot noise, 328, 438
- SI system of units, 54, 280, 364, 419, 421, 431, 432, 441
- silicon, 65, 91, 125, 133, 210, 264, 266
- simultaneity, 185–187, 189, 194, 197, 200, 202–204, 209, 215, 216, 218, 220, 433, 438
- definition, 433
 - loss of, 189, 196, 202, 216, 433
- singularity, 289, 290, 293, 298, 302, 343, 350, 353, 358, 408, 415, 425, 434, 438
- Sirius A and B, 118, 143, 145, 153
- skiing, 12
- slingshot mechanism, 51, 52, 56, 57, 59–61, 63, 155, 234, 380, 429
- slow-time storage locker, 289
- slowing of time, *see* time, dilation; redshift, gravitational
- SN1987A, 129, 130, 149
- Snider, J L, 16
- SNO, *see* Sudbury Neutrino Observatory
- SNU, *see* solar neutrino unit
- SOHO satellite, 85, 98, 102
- solar constant, 106, 108, 438
- solar neutrino unit (SNU), 128, 438
- solar neutrinos, 102, 127–130, 173, 370, 382, 434, 438, *see also* Homestake Gold Mine
- flux, 128–131, 434, 438
- Solar System, 7, 14, 18, 23, 27, 29, 33, 36, 37, 39, 46, 51–54, 56, 57, 59–62, 65, 66, 69, 82, 90, 91, 98, 103, 105, 121, 126, 146, 153–155, 163, 176, 179, 180, 225, 230–232, 235, 236, 238, 241, 257, 260, 264, 266, 309, 318, 329, 339, 347, 359, 395, 404, 412, 422, 424, 425, 429, 440
- formation, 65, 133, 435
- Soldner, *see* von Soldner
- South African Observatory (SAO), 153, 267, 367
- space, 183, 188–190, 203, 210, 213, 214, 216–220, 222, 227, 230, 231, 234, 235, 239, 240, 242, 244, 251, 254, 258, 280, 281, 287, 292, 302, 303, 305, 310, 320, 348, 353, 365, 383–390, 392, 396, 402, 405, 409, 416, 417, 421, 423, 425, 429, 434, 438
- space missions
- communicating with, 329
- Space Telescope Science Institute (STScI), 104, 136, 142, 149, 164, 165–167, 177, 228, 331, 332, 353
- spacecraft, 20, 39, 46, 51, 52, 54, 56–61, 63, 85, 86, 93, 118, 201, 207, 208, 309, 329, 330, 354, 364, 429, 440
- spacelike, 189, 203, 220, 389, 438
- spacetime, 189, 190, 213, 214, 216–225, 229–237, 245, 256–260, 280, 281, 285, 287–289, 295, 304, 305, 317, 336, 337, 339, 360, 389, 390, 407–409, 415, 421, 424, 425, 427–429, 432–434, 438, 441
- diagram, 214–216, 219, 438
 - foam, 295, 303, 408, 409, 416, 438
 - interval, 213, 214, 217–221, 230, 231, 233–237, 242, 244, 249, 251, 258, 280, 285–287, 292, 389, 438, 440
 - negative, 217
 - zero, 219, 220, 287
- metric tensor, 219, 427, 438
- Minkowski, 219, 221, 223, 224, 285, 386, 388, 389, 431, 434
- of special relativity, 214, 219, 221, 223, 224, 232, 320, 386, 388, 424, 431, 434
- spacetime geometry, 213, 219, 232, 234, 241, 249, 257, 317, 339, 424
- flat, 320
- spacetime structure
- elsewhere, 39, 97, 134, 220, 301, 305, 335, 347, 379
 - event, 23, 76, 85, 112, 125–127, 149, 155, 159, 162, 165, 177, 183, 189, 190, 200, 202–204, 210, 213–222, 230, 235, 236, 266, 270, 279, 299, 310, 311, 317, 321–326, 338, 339, 342, 345–347, 350, 354, 382, 389, 406, 407, 411, 424, 427–433, 438, 440, 441
- special relativity, 14, 49, 88, 131, 145, 146, 179–184, 184, 185, 187, 188, 191–196, 198, 202, 205, 207, 209–211, 213, 214, 217, 219–221, 223–225, 231, 232, 234, 236–239, 243, 245, 247, 252, 253, 256, 259, 277, 281, 288, 303, 309, 310, 313, 336,

- 360, 383, 386, 388, 389, 400, 405, 406, 421, 422, 424, 428, 429, 431, 433, 434, 437, 438, 440
 mass depends on speed, *see* mass, dependence on speed
 tests of, 179, 183, 195
 spectral lines, *see* spectrum, lines
 spectroscopic binary, *see* binary stars, spectroscopic
 spectrum, 15, 86, 89, 101, 102, 107, 109–117, 137, 153, 154, 156, 159–161, 168, 172, 175, 267, 271, 277, 300–302, 304, 314, 332, 342, 352, 354–357, 373, 380, 402, 408, 421, 422, 425, 429–431, 434, 438–441
 absorption, 115, 421
 acoustic, 99, 101, 314
 black body, *see* black body, spectrum
 lines, 114, 115, 153, 154, 156, 169, 175, 266, 267, 271, 395, 438
 speed, 2–6, 9, 13–16, 18–24, 26–30, 32, 34–39, 47, 51–62, 70, 73, 74, 77–80, 86, 87, 100, 101, 122, 123, 137, 143, 145, 148, 149, 154–156, 161, 162, 165, 167, 169, 170, 175, 176, 179–184, 184, 185–188, 190–206, 208, 209, 215, 221, 222, 232–234, 246, 248–250, 254, 255, 257, 260, 264–266, 268–270, 275, 277, 279, 280, 287, 290, 291, 293, 297, 310, 317, 318, 320, 323, 339, 343, 348, 349, 352, 355–358, 360, 361, 363–365, 371, 374, 383, 384, 388, 389, 401, 406, 407, 413, 421, 423–425, 427, 428, 430, 432, 433, 437, 438, 440
 speed of light, *see* light, speed
 spherical symmetry, 64, 235, 280, 359, 431
 spin, 45, 46, 66, 142, 144, 206, 207, 252, 261, 266, 268, 270, 273–276, 278, 296, 297, 299, 300, 302, 308, 318, 322, 399, 414, 415, 421, 423, 428, 437, 438, 440
 spiral galaxy, 103, 104, 136, 164, 165, 166, 169, 170, 332, 426, 438, *see also* galaxy
 bulge, 166
 disk, 166, 167
 spontaneous symmetry breaking, 399, 400, 405, 425, 438
 St John River, 39
 stability, 46, 100, 101, 126, 132, 133, 135, 143, 146, 172, 173, 264, 265, 271, 275, 276, 283, 284, 291, 299, 302, 339, 382, 411, 431, 434, 439
 stable, *see* stability
 standard candle, 108, 168, 321, 322, 350, 390, 430, 431, 438, 439
 standard meter stick, 390, 439
 star, 6, 7, 16, 18, 23, 25, 29, 34, 36–38, 46–48, 65, 66, 78, 81, 84–87, 89–98, 101, 103–111, 113, 114, 116–128, 130–151, 153–164, 164, 165–175, 177–180, 185, 193, 194, 201, 206–210, 212, 231–233, 235, 236, 239, 240, 243, 244, 251, 252, 259–267, 267, 268–287, 290–293, 296, 299–301, 309, 310, 313, 314, 316–325, 331–343, 345, 347, 349, 350, 354, 358, 360, 362, 366, 367, 369, 370, 373, 375, 375, 376–378, 393, 394, 396, 404, 405, 410–413, 415, 417, 421–430, 433–435, 437–441
 boson, 143, 338, 339, 423
 center, 141
 cluster, 120, 139, 164, 165, 429, 432, 434, 438, 439
 center, 165
 globular, 158, 162, 164, 164, 165, 166, 171, 322, 428, 429, 432, 439
 degenerate, 143, 146
 early generations, 126
 evolution
 contracting core, 140
 first generation, 106, 124, 126, 322, 350, 372, 436
 formation, 124, 131, 135, 136, 138, 166, 167, 172, 177, 274, 412, 431, 432, 438
 initiating, 166
 regions of, 166
 giant, 25, 61, 65, 76, 91, 95, 103, 107, 118, 121, 125, 134, 135, 139, 140, 147–149, 151, 153, 158, 160, 166, 182, 207, 261–263, 266, 269, 277, 278, 282, 300, 313, 314, 350, 411, 412, 423, 429, 430, 433, 434, 436, 439–441
 blue, 140, 149, 429
 red, 139, 140
 massive, 133, 138, 140, 141, 146, 148, 151, 165, 166, 172, 279, 438
 model of, 436
 Newtonian, 281
 polytrope, 282
 neutron star, *see* neutron star
 protostar, 136, 137, 158, 437
 radius, 141, 161
 relativistic
 spherical, 360
 spectrum, 114, 153, 154
 stability, 137
 stellar wind, 140, 148
 structure, 280, 281
 supergiant, 140, 439
 surface temperature, 139, 141
 temperature, 110, 118, 139, 141
 twinkling, 105
 white dwarf, *see* white dwarf
 star cluster, 139, 439
 starquake, *see* neutron star, starquake
 stars
 multiple, 23, 29
 old, 126, 373
 post-main-sequence, 140
 pre-main-sequence, 139
 relativistic
 supermassive, 175
 stationary limit, *see* black hole, Kerr, ergosphere
 statistical mechanics, 76, 112, 182, 306, 308, 439
 steady-state theory, *see* cosmology, steady-state
 Stefan–Boltzmann constant, 116, 305, 419
 Stefan–Boltzmann law, 116, 117, 119, 356
 stellar model, *see* star, model of
 strain, 71, 312, 439
 strange matter, 284, 439
 strange stars, 284, 439
 stratosphere, 82
 stress, 238, 258, 427, 439
 stress–energy tensor, 258, 439
 string theory, 308, 388, 404, 409, 423, 436, 437, 439, *see also* quantum, gravity; unified field theory
 STSCI, *see* Space Telescope Science Institute
 Stukeley, W, 178

- Sudbury Neutrino Observatory (SNO), 130, 131
- Sun, 6, 7, 12, 13, 16, 18, 20, 23, 25, 27–35, 37, 38, 40, 42, 43, 45–49, 51–68, 70, 80–82, 85, 87–106, 108, 110, 113–115, 117–131, 133, 135, 137–141, 143, 144, 146, 148, 153–159, 161, 168, 169, 172, 174, 176, 177, 185–187, 201, 210, 222, 225, 230, 231, 233, 235–237, 239, 241, 242, 244, 252, 256, 257, 259, 261, 263, 266, 267, 269–273, 275, 276, 280, 281, 283–285, 294, 309, 310, 314, 318, 325, 329, 332, 347, 355–359, 370, 376, 405, 412, 419, 422, 424, 426, 430, 432, 434–436, 438, 439
- center, 37, 85, 88–92, 94–96, 115, 128, 130
- corona, 98, 115, 118
- escape speed, *see* escape speed, Sun
- flux at Earth, *see* solar constant
- heating by, 82
- luminosity, 106, 133, 139, 169, 305, 319
- magnetic field, 98
- mass of, 28, 29, 31, 37, 38, 54, 65, 92, 95, 97, 105, 118, 133, 134, 137, 140, 143–145, 147, 153, 165, 174, 175, 237, 241, 257, 259, 264, 276, 278, 279, 281, 291, 292, 318, 339, 378, 419, 435, 436, 441
- normal modes
- measuring frequencies, 102
- orbit, 33
- due to Jupiter, 48
- radius, 28, 33, 37, 93, 94, 96, 97, 117, 118, 419
- rotation, 98
- seismology, *see* helioseismology
- solar wind, 82, 98, 140, 271, 278
- standard model, 92–95, 102, 128–131
- surface temperature, 115, 300, 376
- temperature, 88, 89, 115
- tidal effect on Earth, *see* tide, due to the Sun
- sunburn, 85–88
- sunspots, 98, 296, 439
- numbers of, 98
- supercluster, 170, 379, 439
- supercomputer, 70, 136, 173, 177, 239, 240, 258, 292, 323, 325, 366, 377
- superconductor, 261, 284, 439, 441
- superfluid, 261, 284, 429, 439–441
- supernova, 103, 123, 125, 126, 128–130, 133, 135, 136, 140, 147, 149–151, 159, 160, 167, 168, 171, 207, 210, 263, 266, 267, 268–270, 273, 274, 284, 310, 314, 324, 326, 331, 350–352, 371, 402, 403, 406, 410–412, 425, 429, 431, 433, 435, 439, 440
- rate, 270
- remnants, 267, 268, 277
- super-supernova, 279
- Type Ia, 151, 352, 403, 439
- Type Ia, 151, 160, 168, 350, 351, 402
- Type II, 149, 151, 429, 431, 440
- Type II, 149–151, 268, 270, 324, 350
- surface brightness, 335, 440
- surface of last scattering, *see* Universe, surface of last scattering
- surface of sphere, 213
- symmetry, 11, 14, 35, 41, 239, 241, 246, 249, 251, 258, 320, 360, 399, 400, 402, 405, 425, 431, 438
- synchronous rotation, 45, 440
- tachyon, 196, 203, 440
- TAMA, *see* interferometric detector, TAMA
- tau meson, 374
- Taylor, J I, 317–319
- telescope, 7, 19, 25, 46, 47, 104–106, 153, 167, 168, 176, 177, 228, 270, 271, 315, 317, 332–336, 339–343, 353, 393
- Tell, W, 214–218, 220, 222
- son, 214
- temperature, 22, 65–71, 75–85, 87–96, 98–101, 107, 111–120, 122, 124, 128, 129, 133, 136–141, 148, 151, 159–161, 171, 194, 201, 261, 263, 265, 267, 277, 278, 280–282, 300, 301, 304–308, 327, 354–357, 365, 368–370, 372, 374–376, 379–382, 384, 391, 393, 398–400, 407, 421–431, 435, 438, 441
- absolute, 75–78, 432, *see also* kelvin
- scale, 75
- absolute zero, 65, 75–77, 80, 83, 84, 432, 441
- inversion, 82
- scale, 75, *see also* celsius, kelvin
- tension, 80, 99, 244, 254, 257, 258, 405, 434, 440
- in rubber band, 244, 254, 434
- tensor, 258, 424, 427, 431, 434, 438–440
- test-ban treaty, 278
- thermal
- equilibrium, 96, 133, 149
 - kinetic energy, 88, 159
 - motions, *see* random, atomic motion and temperature
- thermodynamics, 93, 111, 182, 296, 307, 427, 440
- thermonuclear, *see* hydrogen bomb
- thermosphere, 82
- thermostat, 138
- Thirring, H, 252
- Thorne, K S, 288, 322, 324
- thought experiment, 16, 195, 197, 200, 201, 204, 206, 399
- three-body collision, 66, 165
- three-sphere, 387–390
- tidal
- acceleration, 40, 43, 311, 440
 - force, 21, 39–48, 153, 154, 157–159, 177, 224, 279, 289, 310, 311, 327
 - inverse-cube law, 43
- gravity, 157, 165, 167, 329
- interval, 42
- range, 44
- tide
- arrival time, 44
 - due to Sun, 42
 - due to the Moon, 45
 - neap, 42, 43
 - spring, 42, 43
- time, 3–6, 17, 18, 23, 38, 46, 48, 57, 62–64, 78, 81, 82, 113, 125, 133, 134, 146, 150, 151, 176–179, 181, 183, 188–191, 194, 200, 202, 208–210, 213–221, 230, 231, 234, 235, 237, 243, 244, 254, 256, 258, 260, 274, 285–287, 289–292, 307, 311, 312, 329, 331, 343, 345, 346, 348, 349, 366, 370, 382, 390, 395, 400, 401, 405, 415–417, 419, 422, 426–428, 437, 438
- arrival, 278
- arrow, *see* arrow of time

- beginning of, 120, 179, 345–347, 350, 353, 394, 408
computation, 30
coordinate, 219, 222, 230–232, 251, 286, 287
 rescaling, 219
curvature, *see* curvature, time
decay, 188, 197, *see also* half-life
delay, 12, 42
dilation, 17, 184, 187–189, 191–193, 195, 197–199, 201, 202, 204, 205, 207–209, 215, 216, 218, 221, 235, 289, 291, 303, 389, 406, 415, 422, 428, 438, 440
apparent paradox, 198
appetite-suppressant, 208
 due to gravity, 17, 289
end of, 350, 408
exposure, 270
interval, 3, 4, 23, 30, 38, 53, 54, 58, 78, 126, 161, 191, 213, 216, 230, 234, 250, 276, 356, 370, 389
light travel, 197, 200, 201
light-travel, 16, 38, 186, 215, 219, 276, 295
measuring, 3, 188, 200, 203, 208, 219, 221, 231, 438
not universal, 216
proper, 221, 230–232, 286–288, 291, 293, 437
 stands still for light, 131, 188, 440
psychological, 201, 415, 422
quantum nature of, 417
reversal, 118, 161, 313, 358, 368, 441
 symmetry violation, 412, 416, 422
running backwards, 125, 203, 249, 251, 311, 313, 315, 346, 351, 366, 373, 409, 416, 440
simultaneous, *see* simultaneity
travel, *see* time travel
time travel, 303, 409, 415, 440, 441
time-step, 3–5, 22, 23, 28–31, 56, 58, 126, 157, 162, 364
 adjustment, 30
 halver, 29
timelike, 190, 203, 220, 221, 389, 440, 441
 future, 220
 past, 220
Titan, 82, 83
Tokyo, 328
Tomonaga, S-I, 395
trajectory, 2, 4–6, 20, 23, 24, 28, 37, 38, 53, 56, 57, 59–61, 64, 158, 231–235, 237, 245, 304, 337, 343, 364, 426, 429
transponder, 329, 330, 440
transverse, 38, 204, 311–313, 429, 433, 436, 440
trigonometry, 5, 22, 23, 337
triple-alpha collision, 133
twin paradox, 198, 207
two-hand rule, 248, 249, 251, *see also* right-hand rule
ultraviolet radiation, 68, 82, 85, 87, 88, 95, 109, 435, 440, 441, *see also* electromagnetic waves
uncertainty principle, *see* Heisenberg uncertainty principle
unified field theory, 14, 182, 409, *see also* quantum, gravity; string theory
United States Air Force, 278
United States of America (USA), 150, 269, 307, 327, 328, 423
units, 3, 3, 5, 10, 27, 29, 31, 54, 56, 72, 79, 87, 96, 104, 106, 107, 111, 116, 117, 124, 139, 162, 167, 219–221, 228, 235, 243, 246, 280, 300, 302, 315, 319, 348, 364, 368, 397, 406, 410, 419, 422, 426, 433, 441
Universe, 18, 19, 39, 63, 77, 84, 98, 103, 107, 109, 111, 121, 124–126, 131, 134, 143, 145, 151, 163, 164, 168, 170, 173, 175, 177–180, 183, 192, 193, 210, 212, 217, 241, 251, 253–257, 260, 261, 266, 279, 284, 293, 294, 301, 303, 309, 322, 324–326, 330, 331, 333, 335, 342, 343, 345–375, 375, 376–378, 380–384, 388–395, 397–405, 407, 409–415, 417, 421–426, 428, 430, 431, 434, 435, 437, 439–441, *see also* cosmology
acceleration, 151, 168, 239, 241, 244, 256, 345, 351, 352, 358, 371, 395, 402
age of, 124, 305, 325, 348, 352, 353, 359, 363, 366
center, 365
declaration, 351, 362, 363, 426
decoupling, 354, 355, 376, 377, 380, 382, 401, 412, 426, 430, 432
density, 254, 361, 362, 370, 374, 377, 389, 392, 397, 398, 403, 426
early, 131, 173, 175, 177, 244, 305, 325, 354, 356, 363, 365, 368–373, 375–377, 381, 389–391, 397–400, 402, 409, 416, 421, 425, 431, 433, 436
escape speed, 358
expansion, 63, 107, 126, 132, 167, 168, 173, 212, 239, 253, 256, 348–353, 358, 362–365, 368, 372, 376, 381, 392, 394, 395, 403, 412, 416, 425, 426, 428, 430, 432, 437, 439, 441
 anisotropic, 384
expansion rate, *see* Hubble, constant
future, 358, 388, 393
matter density, 364, 376
observable, 325, 347, 353, 428
surface of last scattering, 96, 357, 436, 440
time-reversed, 367
unstable, *see* instability
uranium, 69, 125, 126, 151, 206, 210, 261, 263
Uranus, 25, 28, 48, 51, 66

v404 CYG, 299
vacuum, 181, 187, 206, 255, 324, 327–329, 398, 399, 401, 405
variable stars, 168
 cataclysmic, 158, 160, 423
 Cepheid, 168
 period, 168
vector, 20–23, 243, 424, 426, 431, 436, 438, 440
 position, 22, 436
Vela satellites, 278
velocity, 17, 20–23, 30, 52, 53, 56, 58, 77, 78, 84, 100, 123, 142, 147, 148, 154, 156, 157, 160, 162, 164–166, 168–171, 175, 191, 192, 212, 222, 233, 241, 242, 245, 246, 248–251, 253, 267, 269, 270, 291, 295, 316, 339, 356, 357, 363, 366, 371, 381, 384, 389, 401, 404, 421, 424–426, 430, 434, 438, 440, 441
absolute, 357
Venus, 7, 25, 28, 51, 58, 60, 61, 66–69, 82, 83, 155, 158, 259, 440
Very Large Array (VLA), 174

- vibration, 84, 99–102, 112, 113, 115, 182, 185, 271, 284, 293, 314, 325–329, 396, 399, 435
 characteristic, *see mode*
 violin string, 99, 396
 VIRGO, *see* interferometric detector, VIRGO
 Virgo cluster, 164, 170, 314, 317, 406, 428
 virial method, 171, 440
 viscosity, 284, 440
 VLA, *see* Very Large Array
 volcanism, 46, 131
 volcano, 25, 46, 66–69
 eruption, 39, 46
 von Soldner, J G, 38
 vortex, 284, 440, 441
 Voyager spacecraft, 60
 Voyager 1, 46
 Voyager 2, 39
 Washington University Relativity Group (WASHU), 239, 292
 water, 26, 31, 41, 44, 67, 68, 70, 72, 75, 82, 88, 89, 123, 124, 144, 149, 187, 196, 250, 251, 260, 261, 293, 311, 317, 371, 422, 423, 440
 heavy, 127, 130
 vapor, 66–68, 70
 water waves, *see* waves, water
 watt (unit for power), 54, 441
 wave-particle duality, 86
 waves, 15, 17, 44, 86, 87, 98, 102, 149, 166, 182, 185, 187, 191, 192, 196, 309–317, 321, 322, 328, 329, 395, 396, 416, 421, 422, 426, 429, 433, 435–438, 440, *see also* electromagnetic waves; gravitational waves; radio waves
 bore, 39, 44, 149, 422
 plane, 316, 436
 seismic, 99
 shock, 149, 151, 166, 266, 411
 sound, 15, 77, 100, 309, 311, 382, 433
 water, 86, 187, 309, 317, 422, 433, 440
 weakly interacting massive particle (WIMP), 173, 381
 weather, 19, 25, 70, 82
 Weber, J, 326, 327
 gravitational wave detection claim, 326
 website (for this book), 5, 23, 28, 35, 37, 38, 56, 81, 82, 93, 94, 97, 119, 155, 157, 162, 280, 282, 290, 364, 403
 weight, 2, 10, 11, 13, 20, 21, 23, 71–73, 75, 79–81, 92, 94, 125, 126, 139, 235, 296, 441
 atomic, *see* atomic weight
 sensation of, 21
 weightless, 2, 17, 20, 21
 Wheeler, J A, 285, 288, 308
 white dwarf, 118, 120, 135, 140, 143–151, 154, 159–161, 165, 262, 264–266, 271, 272, 276–278, 281, 286, 300, 318, 339, 350, 410, 411, 423, 430, 433, 436, 439, 441
 binary, 165
 maximum mass, 145
 white hole, 303, 441
 Wien's law, 117, 119, 201
 Wilson, R W, 355
 WIMP, *see* weakly interacting massive particle
 work, 2, 62, 190, 194, 205, 257, 356, 362, 436, 440, 441
 work function, 86
 world line, 214–216, 219–223, 231, 286, 287, 297, 298, 429, 441
 wormhole, 302–304, 408, 409, 415, 440, 441
 X-rays, 95, 103, 131, 153, 160, 161, 164, 171, 172, 197, 261, 267, 267, 268, 271, 272, 276–279, 285, 293, 297–302, 318, 322, 354, 406, 421, 423, 437, 441, *see also* electromagnetic waves
 binary system, 158, 277, 441
 emission, 171, 172, 270, 274, 299, 301
 observations, 171, 277, 319
 region of spectrum, 111, 161
 soft, 160
 source
 first discovered, 271
 sources in astronomy, 261, 267
 telescopes, 171, 267, 276, 299, 300
 XMM satellite, 301, 302
 XN MON 75, 299
 XN MUS 91, 299
 XN OPH 77, 299
 XN PER 92, 299
 XN SCO 94, 299
 XN VUL 88, 299
 Z⁰ particle, 374
 ZIB, *see* Zuse Institute Berlin
 Zuse Institute Berlin (ZIB), 239, 285, 292, 323
 Zwicky, F, 171, 268, 342