

OXFORD

FOUNDATIONS OF MODERN COSMOLOGY

SECOND EDITION

John F. Hawley Katherine A. Holcomb

Foundations of Modern Cosmology

This page intentionally left blank

Foundations of Modern Cosmology

Second Edition

John F. Hawley and Katherine A. Holcomb

University of Virginia

OXFORD
UNIVERSITY PRESS

OXFORD

UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

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

Oxford New York

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

With offices in

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

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

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

© Oxford University Press, 2005

The moral rights of the authors have been asserted
Database right Oxford University Press (maker)

First published 2005

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system, or transmitted, in any form or by any means,
without the prior permission in writing of Oxford University Press,
or as expressly permitted by law, or under terms agreed with the appropriate
reprographics rights organization. Enquiries concerning reproduction
outside the scope of the above should be sent to the Rights Department,
Oxford University Press, at the address above

You must not circulate this book in any other binding or cover
and you must impose the same condition on any acquirer

British Library Cataloguing in Publication Data

Data available

Library of Congress Cataloging in Publication Data

Data available

Printed in Great Britain
on acid-free paper by
Biddles Ltd., King's Lynn, Norfolk

ISBN 0-19-853096-X (Hbk) 9780198530961

10 9 8 7 6 5 4 3 2 1

To our parents and other teachers

Preface to the Second Edition

Since the publication of the first edition of this text, the field of cosmology has undergone dramatic changes. Einstein's cosmological constant, once relegated to the status of a historical artifact, has reemerged in the form of "dark energy" as a significant dynamical element in the cosmos. The long-standing question of the geometry of the universe has apparently been answered by data from the *Wilkinson Microwave Anisotropy Probe*: the universe is flat. The matter content of the universe is around 30% of the critical value, with the balance supplied by the dark energy. Only a fraction of the matter is composed of baryons. Although the nature of this unknown dark matter remains elusive, new experimental results have ruled out the neutrino, while at the same time establishing that the neutrino does possess a small nonzero rest mass. Extensive new galaxy redshift surveys are providing new data on large-scale clustering that are essentially in agreement with the new consensus. Some cosmologists have even begun to speak of an era of precision cosmology. Even if we have not yet truly reached a time when cosmological parameters can be measured to two significant digits, it was not so long ago when cosmologists were satisfied for measurements to agree within a factor of two.

For many years the instructional style in cosmology was to present the basic ideas behind the perpetually unanswered cosmological questions. Now for the first time we are confronted by cosmological answers, and that can be unsettling to those who are content to catalog and contrast varieties of speculative cosmological models. In this edition we have updated the text to reflect the new consensus and to present some of the exciting new observational results of the last few years. Although it is undoubtedly the case that the current concordance model will continue to be refined, we have chosen to take this model seriously as the current best description of the structure of the universe. This has led to some de-emphasis of alternative models. The original aim of this text remains unchanged, however; namely, to present the *foundations* of modern cosmology.

Every chapter in the text has been revised and updated, and the illustrations have been improved throughout. The overall sequence is essentially unchanged from the first edition, but some modifications have been made to accommodate new material and increase the instructor's flexibility. Chapters 1 through 3 remain focused on the historical roots of cosmology. Chapters 4 and 5 contain background physics and astronomy; most instructors can assign these as supplemental reading. Chapters 6 through 9 present topics in relativity. Of these, Chapter 6 introduces the cosmological principle and the fundamental idea of relativity, while Chapters 7 and 8 focus on special and general relativity. These chapters can be covered quickly if the emphasis is to be primarily on observational cosmology. Chapter 9 presents black holes, a topic which, although perhaps outside the main thrust of cosmology, is always among the most popular in classes. The next section of the book develops the

essential elements of modern cosmology. Chapter 10 introduces the idea of expanding space. Chapter 11 is the most mathematical, as it deals directly with the Friedmann equation and the derivations of some of the important parameters of the universe. Instructors preferring to minimize the mathematical content may wish to tread lightly there. Some of the basic ideas from Chapter 11 that are necessary to understand the implications of cosmological observations are reiterated in Chapter 13. We have rewritten Chapters 12 through 17 to focus on specific topics of modern cosmology; each chapter more or less stands alone. The detailed physics of the early universe is concentrated in Chapter 12. Chapter 13 discusses how the parameters of the universe can be measured and includes some of the latest results and their implications for cosmology. The cosmic background radiation is now covered in its own chapter, Chapter 14. The emphasis is on observational results, including those from the recent *WMAP* mission. The nature of the dark matter and its implications for cosmic structure are presented in Chapter 15. Inflation is covered in Chapter 16. Chapter 17 concludes the book with a brief discussion of quantum cosmology and speculations beyond the limits of current theory.

Several new pedagogical features have been added. Each chapter now has marginal notes that highlight key concepts. Chapter summaries are now provided as well. The key terms are listed at the beginning of each chapter, with definitions given at the end of the chapter. In addition, all key terms are defined in a glossary at the back of the book.

We have attempted to correct all the errors and other inadequacies of the first edition, both those discovered by us and those brought to our attention by helpful colleagues. It is perhaps inevitable that new ones will have been introduced in the process of revision, but we hope that none will be beyond the ability of the instructors or the students to handle. Another concern is that the rapid pace of cosmological discovery will continue over the coming years, and that this edition will become outdated even more quickly than the first. We can at least hope that this will prove to be the case.

*December, 2004
Charlottesville, Virginia*

JOHN F. HAWLEY
KATHERINE A. HOLCOMB

Preface to the First Edition

Recent discoveries in astronomy, especially those made with data collected by satellites such as the *Cosmic Background Explorer* and the *Hubble Space Telescope*, have brought cosmology to the forefront of science. New observations hold out the tantalizing possibility that the solutions to some especially elusive mysteries might be found in the near future. Despite an increase in public interest in black holes and the origins of the universe, however, the unavoidable lack of context with which discoveries are reported prevents most people from understanding the issues, or appreciating the true significance of the new data. Popular books on cosmology abound, but often they present the subject as a series of “just so” stories, since some basic physics is a prerequisite for comprehending how cosmology fits into modern science. The lay reader may well have trouble distinguishing knowledge from speculation, and science from mythology. Furthermore, the popular literature often emphasizes the more exotic aspects of the field, often at the expense of the firmly grounded achievements of modern cosmology.

Cosmology holds an intrinsic interest for many college students, who are granted, as part of their general education, the time and opportunity to learn more about the scientific discoveries they see described in newspapers and magazines. Most colleges and universities offer a comprehensive introductory astronomy course, with the primary objective of offering science to as broad a population of students as possible. Topics such as relativity, black holes, and the expanding universe are typically of particular interest, but they are covered in a cursory fashion in most introductory courses and texts. In our experience, there is always a sizable number of students who find astronomy sufficiently interesting that they wish to continue their study of the subject at a comparable technical level, but with greater depth. With little but astronomy-major or graduate-level courses available, however, such students often have no such opportunities. These students, who are genuinely interested in learning more about these topics, deserve the opportunity to further their learning, and to do so in a serious way.

The course from which this book grew is intended for upper-division liberal arts students at the University of Virginia. Most of the students who take it have some basic science background, such as would be provided by a general introductory astronomy course; however, well-prepared students can and do take the course in lieu of general astronomy. Students from wide-ranging areas of study have taken this course. Their relative success is not necessarily correlated with their major. Some exceptionally strong students have come from the ranks of history and philosophy majors, while occasionally an engineering or astronomy major has floundered. Extensive experience with math and science are not prerequisites; interest and willingness to think are.

This text is intended to fill the gap between the many popular-level books which present cosmology in a superficial manner, or which emphasize the esoteric at the expense of the basic, and the advanced texts

intended for readers with strong backgrounds in physics and mathematics. The book is self-contained, appropriate for a one-semester course, and designed to be easily accessible to anyone with a grasp of elementary algebra. Our goal is to present sufficient qualitative and quantitative information to lead the student to a firm understanding of the foundations of modern theories of cosmology and relativity, while learning about aspects of basic physics in the bargain.

The level of mathematical detail is always of concern for instructors of undergraduate astronomy. We have aimed for a middle ground; some may regret the lack of calculus and accompanying derivations, while others may recoil from the appearance of any equation. The real difficulty with a topic like cosmology is not the mathematics per se, but the challenging concepts and the nonintuitive way of thinking required. However, without some understanding of the mathematical basis for cosmology, the student may find it difficult to distinguish from mythology; without data and quantitative analysis, science becomes just another narrative. Thus, while we have tried to keep the level of mathematics consistent with minimum college-level algebra, we have not shied from including some equations within the text, rather than relegating them to an appendix or omitting them altogether. The resulting level is comparable to some of the more comprehensive introductory astronomy texts. Of course, more or less mathematical detail may be included or required by the instructor, depending upon the backgrounds and wishes of the students.

The book contains more material than can usually be presented in one semester. The instructor has a good deal of flexibility in designing a particular course. Depending on the background of the students, various sections can be given more or less emphasis.

The text is divided into five major sections. Since many students are unaware of the historical background from which modern cosmology grew, we begin with an overview of historical cosmology, from ancient myths to present scientific theories. The history of cosmological thought demonstrates that the universe is not only knowable to the human mind, but that the modern physical universe, constructed in the light of our new understanding of physics, is far grander than the constricted heavens of the ancients. This section also lays out the important cosmological questions, and introduces the ideas of natural motion, symmetry, and the relation of physical law to the structure of the universe. For students who have just completed a typical introductory astronomy course, the historical and review sections could be covered quickly, with an emphasis on Newton's laws.

The second section exists primarily to make the book self-contained; it quickly reviews points that are likely to have been covered in an introductory astronomy or physics course. We do not assume or require introductory astronomy as a prerequisite; a motivated reader can find all the necessary background material here. While this section can be discussed briefly, or skipped entirely, even those students who have previously studied astronomy might find it beneficial to review this section.

The theories of special and general relativity are presented in the third section, with emphasis on the fundamental physical consequences of these theories. Many textbooks, particularly at the graduate level, de-emphasize relativity, since it is true that little knowledge of the theory is required for the study of cosmology. However, readers who form the intended audience of this text often find relativity particularly fascinating, since it is so drastically different from anything they have previously learned or thought. Relativity is the setting upon which much of modern cosmology takes place, but professional astronomers often take this worldview so much for granted that they do not appreciate the point of view of students who have never encountered this material. Class surveys have consistently shown that relativity makes the greatest impression upon most of the students. In any case, portions of this section are indispensable. Chapter 6 presents the Cosmological Principle, a concept that is obviously required for the remainder of the book. Chapter 7 introduces the essential concepts, including the space-time interval, lightcones, and the metric. Chapter 8 on general relativity includes the necessary introduction to the non-Euclidean isotropic and homogeneous geometries. General relativity is highlighted by a chapter on black holes (Chapter 9), which includes some of the latest astronomical ideas and discoveries. While this chapter on black holes can be omitted, students often find that topic to be the most interesting of all.

The theory of relativity provides the background for the next section, which presents basic modern cosmology. Chapter 10 discusses the discovery of the external galaxies and the expanding universe, and the theoretical interpretation in terms of Einstein's theory of relativity. This leads into Chapter 11, which presents the simplest mathematical models of the universe itself, and the standard big bang models. Chapter 12 deals with the discovery and interpretation of the cosmic background radiation, as well as other modern cosmological observations. The history of the universe, starting from this "bang," follows as the next topic. Throughout, emphasis is given to the standard models, with some discussion of the most likely variants.

The final section covers topics that are the subject of current ongoing research. In this section, we emphasize that the standard model of cosmology has been spectacularly successful as a scientific theory; it simply does not yet provide all the answers. We consider the possibility of dark matter in the universe and the formation of large-scale structure in Chapter 14. Inflationary models have been advanced as a possible solution to some of the quandaries of the big bang; they are presented in Chapter 15 with an explanation of how they might answer these questions. We end in Chapter 16 with the most speculative topics; the unification of the two great triumphs of twentieth-century physics, gravitation and quantum mechanics, as well as the enigma of the arrow of time, time travel, and the fate of the universe itself. Any of the chapters in this final section can be used independent of the others, as time permits. Instructors may wish to supplement this material with

additional information from current research, or from their own notes, as appropriate.

As an aid to the students, each chapter includes a list of key terms and review questions. A glossary of terms is provided at the back of the book. A brief description of scientific notation, units, and physical constants is given in the appendices.

We wish to acknowledge those colleagues and friends who provided comments, criticism and advice during the preparation of this book. We thank Steven Balbus, Jane Charlton, Marc Davis, Dorothy James, Hannu Kurki-Suonio, Karen Kwitter, Michael Norman, Christopher Palma, James Stone, John Tonry, David Weinberg, Mark Whittle, and the many students from Astronomy 348.

*June, 1997
Charlottesville, Virginia*

JOHN F. HAWLEY
KATHERINE A. HOLCOMB

This page intentionally left blank

Contents

Part I History	1
1 In the Beginning	3
2 Cosmology Becomes a Science	25
3 Newton's Machine	57
Part II Background	83
4 Lighting the Worlds	85
5 The Lives of the Stars	119
Part III Relativity	151
6 Infinite Space and Absolute Time	153
7 The Special Theory of Relativity	181
8 The General Theory of Relativity	213
9 Black Holes	245
Part IV The Big Bang	275
10 The Expanding Universe	277
11 Modeling the Universe	313
12 The Early Universe	337
13 Testing the Models	379
14 A Message from the Big Bang	407

Part V The Continuing Quest	433
15 Dark Matter and Large-Scale Structure	435
16 The Inflationary Universe	463
17 The Edge of Time	487
Appendix A: Scientific Notation	513
Appendix B: Units	515
Appendix C: Physical and Astronomical Constants	519
Glossary	521
Bibliography	537
Index	543

Part I

History

This page intentionally left blank

In the Beginning

1

Key Terms:

The gods did not reveal from the beginning, all things to us, but in the course of time, through seeking, men find that which is better. But as for certain truth, no man has known it. Nor will he know it.

Xenophanes (6th century BCE)

On a clear, moonless night, in a field far from city lights, the sky might be the cabinet of some celestial jeweler, displaying glittering points of light on a field of black velvet. A faint, irregular band meanders overhead, like a river of cosmic milk. On any particular night, noticeably bright stars might stand out among the others; on subsequent evenings, an observant watcher would find that these wandering lights had shifted their positions against the backdrop of stars. As the seasons change, so does the sky; some groups of stars visible in summer disappear during the winter, whereas others remain above the horizon all year. In the morning, the Sun appears on the eastern horizon. It climbs upward into the sky, then descends and vanishes beneath the western horizon. As the Sun disappears the stars rise, retracing the Sun's motion from east to west across the sky. The Moon rises as well but keeps its own schedule, independent of the stars. At times, the Moon appears as a silvery disk marked with gray splotches; the imaginative may see a man, a rabbit, or even a beetle in the face of the Moon. At other times, the Moon shows us a crescent, or half its disk. Sometimes, it never appears at all.

Today most people pay little attention to the sky, its contents, and its motions. Electric lights and mechanical clocks have dethroned the celestial sphere from its historical importance in human affairs. The inhabitants of brightly lit cities may never have even seen the stars clearly, much less tracked the motions of the planets. Some would also argue that modern science has removed the wonder from the sky; the planets, the Sun, the Moon, and the stars have all been explained. Yet how many among us understand what those explanations are or what they mean? Romantics often declare that understanding a phenomenon somehow takes away its beauty, reducing it to a desiccated specimen, like a stuffed bird in a museum case. But it is not the scientific understanding that is at fault. The failure to observe, and to ask the questions posed by the observations, shows that the beauty was never truly appreciated in the first place. To those who take the time to look, the sky is still a

The night sky is a source of wonder

marvel, and its wonder is only magnified by the extraordinary discoveries of modern astronomy. The Milky Way retains its grandeur, but now we know that this faint, diffuse light is the combined glow of some of the billions of stars that fill the unimaginably huge galaxy in which we live, an awesome contemplation on a dark evening.

Cosmological questions

The heavens still pose many questions to those who take the time to ponder them. What are the stars? Where are they? What makes the Sun rise, and what carries it across the sky on its daily journey? Where does it go at night, and where are the stars during the day? Why do the wanderers roam among unshifting stars? Such questions follow immediately from even casual observations. From there, the study of the cosmos leads us toward even more profound mysteries. How did it all begin? Was there a beginning at all, or have the heavens and the Earth existed forever? Will the universe come to an end? What is the nature of the universe, and what role might humans play in it?

Many of these questions puzzled the ancients and have long since been resolved; but for the modern observer of the night sky, astronomy has deepened some of the old mysteries and added new ones. Many literate persons have heard such expressions as “the big bang” and “expanding space.” They may be aware that astronomers debate whether the universe is open or closed, infinite or finite, eternal or doomed. But what does it mean to say that the universe expands? Is the universe really expanding? Into what? When astronomers say that most of the mass of the universe is missing, what do they mean? Where could it have gone? What are space and time, and why does time move in only one direction? What is the big bang? How did elements originate? What happens to stars when they die? What is a black hole? What will be the ultimate fate of our Sun, and even of the universe itself? Were there other universes before this one, and will others follow ours?

Cosmology defined

Questions such as these fall within the domain of **cosmology**, the study of the universe. Today we regard cosmology as a modern science, but cosmological yearnings have been part of humanity throughout history. All cultures have a cosmology, for such questions have been asked by all peoples for as long as we have wondered at the stars. The explanations have varied from culture to culture, and from era to era, but all seek to impose an order upon the cosmos, to make it accessible to the human mind. This is just as true of scientific as of prescientific cosmologies, but there is an important difference between the two. Prescientific cosmologies generally interpret the universe in strictly human terms. Early cosmologies certainly began with basic observations; the connection between the changes in the skies and the days and seasons is difficult to miss. Mythological models of the universe sought to render such observations intelligible and to fit them into a theory of existence. However, in the mythological worldview, observations were, for the most part, of secondary importance. Scientific cosmologies, in contrast, are based upon and judged by data, the measurements obtained by direct, objective observations of the universe. The better the data, the better the cosmologies we can develop.

Cosmology can lay a defensible claim to the title “the grandest science,” for no other field can have so vast an object of study: the universe in its entirety. But what do we mean by the universe? We might define the universe as “the sum of all that exists” but this is insufficient, for existence draws its meaning from the universe. The universe exists independent of any, or all, of its contents. A complete definition of the universe may not be possible, for it may be that some aspects of the cosmos are forever beyond our limited understanding. Here we will define the **universe** as that which contains and subsumes all the laws of nature and everything subject to these laws; that is, all that is physical. Is cosmology, then, the study of everything? Are all sciences cosmology? Such a definition would be too broad to be useful. We restrict our definition of cosmology to the study of the formation, structure, and evolution of the universe as a whole. This will prove more than sufficient.

The universe is defined as all that is physical

Cosmology is sometimes regarded as a subfield of **astronomy**, but this is not an accurate division. Astronomy is the study of the contents of the universe. Modern cosmology is intimately linked with astronomy, since the only way in which we can observe the universe is to observe the objects it contains; but cosmology is also closely tied to physics. The universe consists not only of bodies, but also of forces and laws that govern their interactions. Indeed, we shall find that physics plays a much greater role than astronomy in describing the earliest moments of the universe. Cosmology draws upon many fields, and itself contributes to other sciences, sometimes in unexpected ways.

At the dawn of the 21st century, cosmology can take pride in its accomplishments. A coherent view of the universe has emerged, the hot big bang model, which successfully explains a remarkably broad range of observations. While the big bang model has never claimed to represent the final truth, it nonetheless provides a framework for understanding the cosmos from the earliest few fractions of a second of its existence till the present; it even predicts how it all might end. Surely this must count among the greatest of human achievements, even though this model cannot yet explain all of the unknown. In this text we will present the foundations of modern cosmology. The history of the development of scientific cosmology shows keenly how our intuitions and common sense continually mislead us, from our perception of an unmoving Earth, to our persistent belief in the absoluteness and inflexibility of space and time. Nevertheless, we can transcend these human limitations and arrive at a picture of the universe that is much closer to what it truly is.

Cosmology, a human endeavor

Cosmological roots: mythology

Although modern cosmology is scientific, and is based upon highly detailed observations of great sensitivity and precision, the big bang model has a long lineage of human explanations of the cosmos. Most of these ancestral models have much to do with human hopes, desires, and preoccupations, and precious little to do with observations. To some extent,

Anthropomorphic universes

this was due to the limited capabilities of the unaided human senses. Much more important, however, were the philosophical prejudices that prevailed for millennia. Only slowly have humans learned to understand what our senses tell us.

Young children tend to interpret their worlds in terms of themselves; so it was with humanity for most of its history. The earliest cosmologies were anthropomorphic cosmologies. An **anthropomorphism** is the interpretation of that which is not human or personal in terms of human or personal characteristics. Attributing human motivations and emotions to the cosmos as a whole, or phrasing the existence of the universe in terms of a literal birth and death, are examples of anthropomorphisms. One form of an anthropomorphic worldview is *animism*, the belief that all things are animated by spirits, all of which hold some opinion toward humans, and any of which may actively aid or frustrate human plans. Less purely anthropomorphic cosmologies may hold that some portions of the universe are inanimate but are created and affected by animate beings, perhaps by a pantheon of gods as well as humans.

Mythological cosmology

The tendency to anthropomorphism comes from the quite natural inclination of human cultures to describe the universe in terms of imagery from familiar, and necessarily human, experiences. When a cosmology is expressed in the form of a narrative tale that explains or illustrates the beliefs of a culture, it is said to be a **myth**. Cosmological myths make the culture's ideas of the origin and structure of the universe generally intelligible and broadly accessible. Some of these myths were interpreted quite literally and anthropomorphically, while others were understood to be only analogies that could make the incomprehensible more familiar. Mythology tends to reflect what is known or important to the culture in which it arose. The myths of agricultural societies typically revolve around the imagery of the seasons, of planting and harvesting, while the myths of hunting and gathering peoples often involve animals that take on human characteristics.

Ancient mythologies still hold so much power that even modern cosmologists sometimes inappropriately blend objective data and mythological leanings. This is illustrated by the special fascination that cosmological beginnings and endings continue to hold. There is no particular reason to believe that the universe must have a beginning or an end, based solely upon our immediate observations; on the scale of human life, the Earth and sky seem eternal and unchanging. Yet it is also true that seasons begin and end, plants sprout and wither, and animals and humans are born and die. Perhaps, then, the universe too has a beginning and an end. Not all mythologies assume this, nor do all modern scientific models. Even among modern scientists, preferences for one model over another have sometimes been based more upon philosophical beliefs than upon data. A distaste for the big bang or for an infinite expansion is an emotional choice based upon a personal mythology. When the big bang model was first introduced, many of the most prominent scientists of the day reacted quite negatively; such an abrupt beginning for the cosmos was uncomfortable for some of the older generation. Still

others interpreted the big bang as scientific vindication for the existence of a creator. Today the big bang is well accepted on its own objective merits, but now some discussions of the possible end of the universe carry an echo of the aesthetics of the cosmologist. Regardless of such intrusions of human wishes, the major difference between science and mythology stands: in science, the cumulative evidence of data must be the final arbiter.

The recognition of familiar ideas and concepts makes even a cursory study of the mythologies of many cultures an enjoyable pursuit. It is worth remembering that many of our unexamined cosmological ideas, including some most firmly embedded within the human psyche, have mythological origins. Many genesis myths can be seen to share common themes. Three categories of imagery are commonly invoked to explain the beginning of the universe. One is the action of a supreme craftsman, mirroring the image of a human artisan at work. Another is generation from a seed or egg, reflecting biological generation. The third is the imposition of order onto chaos, as in the development of human society. These three are not mutually incompatible, and many myths incorporate two or more of these motifs. Nor are these the only possible themes; the early Hindu creation epic, the *Rig-Veda*, makes no explicit claims about the creation of the universe, suggesting only that perhaps some highest god knows, and hinting that it is beyond mortal comprehension.

Another recurrent theme of great importance to humans, though of less significance to the universe as a whole, is the origin of imperfection in the human condition. Many cultures have believed that humans were originally close to the gods, but sinned and were punished. The origin of death is often attributed to human misbehavior; in a number of traditions, women take the brunt of the blame. Death is not always a punishment, however; some see death as the result of an active choice, such as the choice to be able to have children. (From a biological perspective, this is rather accurate.) Such myths seek to understand humanity's place within the universe, an issue with which we still struggle today.

A few specific examples will serve here; they are by no means intended to be comprehensive, but are fairly typical of the range of themes in cosmological mythology. The myths of a society spring from that society, and these examples vividly illustrate this, but myths, once established, can also mold societies long after they cease to function literally.

The *Enuma Elish* is the "Babylonian Genesis." Babylon, a great city of ancient Mesopotamia located near present-day Baghdad, was originally settled by the Sumerians around 3500 BCE. The Sumerians irrigated the desert and developed the cuneiform writing system, but they were conquered by Akkadians and Amorites from the north, and their culture was assimilated and eventually forgotten. The great king Hammurabi, famous as the first ruler known to have written down a code of laws, was an Amorite ruler of Babylon in the 18th century BCE, during the height of Amorite power in the region. A few centuries later, however, Babylon and its possessions came under the control of the

Creation imagery

The Babylonian Genesis

Kassites, a tribe that may have originated in central Asia. The *Enuma Elish* dates from the Kassite regime, perhaps around 1450 BCE; however, only later copies of it have survived. The second millennium BCE was a peak period in Babylon's history, and this creation myth was probably composed at least partially with the motive of justifying the city-state's political power by making its patron deity the chief among the gods.

The story incorporates ancient Sumerian themes, as well as contributions from the later conquerors of Mesopotamia; like many Babylonian myths, it is evocative of later stories indigenous to the Middle East. In this tradition, the tumultuous sea is identified with disorder. The Sumerians and their successors believed that the cosmos began with a chaos of fresh water, sea, and mist. From this confusion, pairs of gods were created representing the silt, the horizon, and the sky, as well as embodying male and female aspects; this echoes the creation of new land in the delta region between the Tigris and Euphrates rivers, in what is now Iraq. Following the initial creation, there was conflict between order and evil in the form of two deities: Marduk, the protector of Babylon, and Tiamat, the sea-goddess, representing chaos. Marduk killed Tiamat and created the Earth from her body, then created humans from the blood of another rebel god, Kingu. The struggle between two powerful deities mirrors the development and nature of human societies; both good and evil are present, while custom and authority create order, backed if necessary by the application of force.

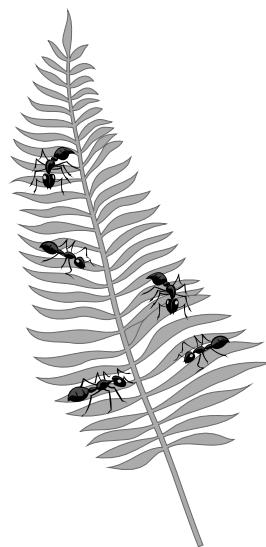


Fig. 1.1 The Word: a Tanzanian myth.

A Tanzanian myth, although quite different in detail from the *Enuma Elish*, similarly reflects the lives of the people who created it. In the beginning was “the Word,” which was the creative force; there were also air and sky, a single Tree, and some ants who lived on the Tree. One day a great wind blew away a branch of the Tree, carrying some ants with it. The ants continued to eat from the branch, but soon they ran out of food and were forced to eat their own excrement. The excrement grew into a huge ball, which became the Earth. The Earth eventually enveloped the Tree, at which point the Word sent wind and water to Earth; the ants were subsequently destroyed in a flood. But the tree continued to grow, and its roots gave rise to plants on the Earth. The atmosphere then created animals and humans, each kind with its own voice. Fighting over food led to war between humans and animals. The war became so terrible that parts of the Earth broke away to form the stars, the Moon, and the Sun. The Sun glows because it came from a part of the Earth that was on fire when it was separated, while the Moon and stars are transparent disks through which the Sun’s light shines. Some of the animals became the slaves of humans, while others remained wild and attacked people. At the end of the war a sheep kept by humans leapt to the sky, where it killed the Word and ruled the cosmos, bringing thunder and lightning to the Earth. Because of this transgression, humans were punished by the gods when they dared to ask for help after their sheep caused the death of the Word. Humans were made lowly and warned that Earth shall be eventually consumed by fire.

Chinese mythology

The best-preserved records from early Asian cultures are those of the Chinese. The Chinese believed that the universe was huge, possibly infinite; the Sun, Moon, planets, and stars consisted of vapor and were blown about by a great wind. Many Chinese also accepted that the Earth moved, though like most peoples they did not realize that it rotated; they envisioned, rather, a smooth oscillation they believed to cause the seasons. Chinese cosmologies emphasized that the doings of humans, particularly the mandarins of the court, were reflected in the heavens, a preoccupation not surprising in a country that was highly and hierarchically organized from very ancient times. Evil works would show themselves by disruptions in the sky, so the Chinese were keen observers, seeking auguries in the stars. Because of this, and because of the antiquity of the Chinese writing system, the Chinese annals constitute the longest unbroken records of the sky, a fact which has proven important for some aspects of modern astronomical research. For example, the Chinese recorded in detail the supernova of 1054; this “guest star,” as they called it, should have been sufficiently bright for a short while to be visible in daylight, yet it is completely absent from European chronicles. Perhaps the Europeans of the time, with their ironclad belief in a perfect, immutable heaven, ignored this strange phenomenon.

A fairly widespread Chinese creation myth tells of the giant Pan Gu. In the beginning, the cosmos was a great egg. For 18,000 years, Pan Gu slept within the egg. Finally he awoke and broke free, shattering the cosmic egg that had contained him. The lighter, purer elements rose and became the heavens, while the heavier, impure elements sank to form the Earth. Pan Gu maintained the separation of Earth and heaven with his body, supporting heaven with his head while his feet rested on the Earth. As the distance between heaven and Earth increased, Pan Gu grew to equal it. Finally heaven and Earth seemed securely in place, and Pan Gu died. His breath became the wind, his voice the thunder, and his perspiration the rain; his left eye was transformed into the Sun and his right eye into the Moon. His four limbs became the four directions, his trunk the mountains, while his blood ran as the rivers and his veins laid out roads and paths. His flesh created fields and soil, his skin and hair became the plants of the Earth, while his bones went into rocks and his marrow became the precious gems. After the sacrifice of Pan Gu, the Earth was a pleasing place, but the goddess Nu Wa, who had the face of a human but the body of a dragon, found it lonely. Stooping by the bank of a pond, she fashioned some amusing little creatures from mud. It was too tiring to create them constantly, so she endowed them with marriage and the capability to reproduce on their own. Later, a great battle between the spirit of water and the spirit of fire resulted in a catastrophe when the fleeing spirit of fire struck the great mountain that supported the western part of the sky. The heavens tilted and ripped apart, while the Earth fissured. Nu Wa melted the prettiest stones from the riverbeds to repair the holes and cracks, then killed a giant turtle and cut off his legs to form the four pillars that support the sky. But

Mesoamerican mythology

the tilt to the west remained and thus the Sun and stars slide down it, while on the Earth the water runs eastward into the ocean.

A prevalent theme among many peoples indigenous to the Americas, both North and South, is the primacy of the four directions and of the number four in general. In some Native American cosmologies these directions correspond to the four cardinal directions also utilized by Europeans, whereas in others the major directions are those of the rising and setting of the Sun at summer and winter solstices. Some Native American cosmologies add the center to the directions, making five the principal mystic number. In many American cultures, particularly in the Southwest of the United States and in Mesoamerica, the world has been destroyed and recreated four or five times. Each world consists of layers, typically three: an Upper World of spirits and pure birds, a Middle World of humans and animals, and a Lower World of evil creatures.

The most elaborate cosmologies of this kind were found in Mesoamerica; the best-preserved version is associated with the dominant Aztec and Mayan tribes. In their belief, the cosmos passed through four ages in the past. At the end of each Sun a great disaster destroyed the world. The current era, the Fifth Sun, began with the self-sacrifice by fire of Nanahuatl, the ugliest god, who was reborn as the Sun Tonatiuh. After this the god Teucciztlan, whose courage had faltered at the great bonfire, threw himself into the flames and became the Moon. But the new Sun sullenly refused to rise until it was placated with the sacrifice of hearts and blood. Xolotl, a twin and aspect of the serpent god Quetzalcoatl, performed the sacrifices of all 1600 deities present, then sacrificed himself; after this, the Sun rose. Quetzalcoatl (the plumed serpent) was, in many Mesoamerican traditions, a creator god who took several aspects, including the wind god Ehecatl and the monster Xolotl. He was a very important deity in Mesoamerica, especially to the Maya. They called him Kukulkan and associated him with the planet Venus, which as the Morning Star rises in the east just before the Sun, and as the Evening Star was thought to plunge sacrificially into the Sun just after it sets. Mayan astronomers kept meticulous records of Venus, and their observations enabled them to compute the length of the solar year to within a few seconds.



Fig. 1.2 El Caracol temple in Mexico. Built by the Mayans around AD 1000, this temple was used as an astronomical observatory to record such celestial events as the rising and setting of Venus.

The importance of Quetzalcoatl played a pivotal role in the conquest of Mexico by the Spanish. Legends told of a great battle between Quetzalcoatl and his rival, the jaguar god Tezcatlipoca, after which Quetzalcoatl disappeared, promising to return from the east. The year 1519 corresponded roughly with the year Ce Acatl (One Reed) in the Aztec-Mayan calendar, the date-name associated with Quetzalcoatl as the Morning Star. When Hernán Cortés and his men appeared on the east coast of Mexico, the Aztec emperor Montezuma II took him for a representative of the returning Quetzalcoatl, and sent treasures of gold and silver from his capital of Tenochtitlán. The riches merely whetted the Spaniards' appetites for conquest, and they quickly made alliances with tribes held vassal by the Aztecs. The Conquistadores, hardly the salt

of their own society, soon enough demonstrated by their behavior that they were not gods, but merely an unfamiliar and especially rapacious kind of human; yet, strangely, Montezuma persisted in his delusion until Cortez appeared at Tenochtitlán and threw the pious emperor into prison. Montezuma was stoned by his own people for his failure to resist the invaders, and he died a few days later.

The example of Montezuma should make it apparent that cosmological considerations are not idle speculations, but can have significant consequences for the individual and society. Creation myths reflect the values and observations of the cultures that created them. Culture shapes the worldview of its society, and conversely. The actions of the society's leaders, for good or ill, can be dictated by the prevailing cosmological mythology. Even in our modern, industrialized societies, many unspoken cosmological assumptions mold our thinking. One of the most significant is the belief that the bounty of the universe is without limit. Though rarely articulated explicitly, this principle pervades many cultures, encapsulating the view that resources and opportunities are infinite. This point of view fits nicely with the attitude that the Earth is here for the benefit of humanity. As has become increasingly apparent since the middle of the 20th century, such an outlook has important, and perhaps disastrous, consequences. Much of current economic theory is founded upon the postulate that growth can continue indefinitely; that if we run out of some resource, a substitute can always be found. Yet it is clear that the illusion of boundless resources occurs only because the Earth is much larger than a human being, and geological timescales are much longer than a human lifespan. Our perceptions of the Earth, its history, and its contents, are skewed by our human limitations.

The perception of bounty continues to affect modern thought, sometimes in unexpected ways. Even those who recognize the limitations of Earthly resources often argue that space exploration and colonization can provide the materials and living space for a human population that grows without bound. Since the dawn of civilization, the human population has grown at an exponential rate; that is, at a rate for which the increase in population is always approximately proportional to the current size of the population. But if humanity continued to reproduce at such a pace, eventually expansion into space could not occur fast enough to accommodate the new population. Indeed, in a relatively short time, by astronomical standards, we would reach the point at which all the particles in the observable universe would be required just to make up the physical bodies of people. Obviously, this is absurd. Nature will take care of our numbers, by its own methods, if we choose not to do so ourselves. As we achieve greater control over our immediate environment, we require an increasingly better assessment of how we fit into the greater world. We may be just as self-assured as Montezuma, and ultimately just as surprised when we find that the way the world *is*, is quite different from how we *believe* it to be.

Mythological cosmology can have real consequences

The “I” in the center of the universe

Anthropocentrism asserts that humans are central

Mythology casts ideas and aspects of the universe into human terms. In some respects, this is essential to our comprehension; we can deal with such issues only in terms we can understand, which must, by necessity, be of human construction. We are mistaken, however, if we invert this assertion, and assume that humanity is essential to the universe itself. Yet the attitude that humankind occupies a special place in the universe is an overriding theme in almost all mythology. This is **anthropocentrism**, the belief that humans are important to the universe, which may well have been created especially for their purposes. To early peoples, observation seemed to support this viewpoint. The Earth is big, while the Sun and planets and stars seem small. All celestial objects appear to revolve around the Earth. Humans have power over plants and animals. The Earth provides the things that make human life possible, so it must have been created for us. (Early peoples did not generally consider the obvious alternative, that humans require for life what the Earth was able to provide. That is, humans are adapted to the Earth, rather than the Earth being designed for humans.) In contrast, some phenomena, such as the weather, remain beyond our power. These things are important, both blessings and curses to humans. For instance, weather brings rain for crops, but also storms that destroy. Since anthropocentric cosmologies assume that humans are cardinal, these natural powers demonstrate that a still greater power exists, which is inflicting upon us the good and the bad; if we may not be in charge, at least we occupy much of the attention of the powers that are. The aspect of punishment is often central, sometimes almost an obsession; humans did wrong and were punished, hence bearing forever the responsibility for death, decay, and imperfection.

Anthropocentrism is still a powerful concept in popular thought. The most familiar of the many possible examples may well be astrology—the belief that the planets and stars themselves relate to personal actions and destiny. Astrology is one of the oldest systems of belief known. The version that is common in Western countries is based upon a systematization of ancient lore by the Greek scholar Ptolemy, whose *Almagest* still forms the basis of the casting of horoscopes. Astrology is based upon the supposition that the stars influence our lives in mysterious ways, or foretell our destinies through their motions and configurations. Before there was any understanding of gravity or of the orbits of planets, some explanation had to be devised for the regularity of the celestial motions. In the prevalent anthropocentric view, those motions must surely have something to do with human events. In Greek and Roman belief, the planets were explicitly associated with specific gods and goddesses, whose names they still bear. The five planets known to the ancients, those that are visible with the unaided eye, are Mercury, Venus, Mars, Jupiter, and Saturn. The Sun and Moon were also considered planets, making seven in all. The days of the week correspond to these seven planets. Sunday is the Sun’s day, Monday the Moon’s, Tuesday is ruled

by Mars.¹ Wednesday is the day of Mercury, Thursday is governed by Jupiter, Friday corresponds to Venus, while Saturday belongs to Saturn.

The gods and goddesses of ancient Rome may have faded to amusing anachronisms, but astrology still holds the attention of many people. Who has not had the experience of reading the appropriate horoscope in the newspaper and finding that it applies perfectly? This is an example of a phenomenon well known to psychologists. People are much more likely to believe very general statements about themselves, than they are to accept genuine specific psychological assessments. Moreover, there is the universal tendency to interpret vague descriptions in terms appropriate to the individual reading them. Finally, there is the phenomenon of *selective memory*, in which hits are remembered vividly, while misses are forgotten. Even if astrology had never been developed, it seems likely that people would be drawn to some similar system, such as paranormal phenomena, unidentified flying objects, channeling of spirits, past lives, and so forth. Many humans are unwilling to believe that their lives are subject to random occurrences; the wish to seek order in the cosmos is powerful.

Astrology may be easy to ridicule, but other common viewpoints are no less anthropocentric. For example, many believe that the land, the sea, the air, and the animals and plants exist primarily for our benefit, to be used as we see fit. Even if we do not believe in astrology *per se*, we frequently believe that we must deserve our fates; our goodness or badness determines the vicissitudes that befall us in life. We believe in cause and effect, but even more, we have a strong desire to believe that the causes of events are purposeful, not due to chance. If they are purposeful, they are understandable, predictable, and controllable. However, if the behavior of the universe were controlled or dictated by the needs and actions of some 6 billion humans, with their conflicting motives and desires, then we might as well return to the ancient myths of unpredictable gods.

The triumph of scientific cosmologies over the anthropocentric world-views has not always been welcome; many people mourn the ancient universes in which humans played a clear and important role. The new universe seems, to some, a bleak and sterile place, while the ineffable universes of the past seemed awesome and meaningful. But this attitude often results from a confusion of the knowledge of a thing, or, more precisely, the model that allows us to know it better, with the thing itself. Science knows that crystals are highly ordered arrangements of atoms; quartz, for example, is simply a chunk of a common mineral, a major component of sand, which happens to have an ordered structure. It is the unusual large-scale symmetry of crystals, compared to most objects, that accounts not only for their rarity in nature, but also for

Modern examples of anthropocentric beliefs

¹In English, most of the names of the days of the week come from Norse gods and goddesses who played roles similar to those of the Graeco-Roman deities.

their beauty.² But this leaves many people dissatisfied; they feel that the ability of polished crystals to refract light, which sometimes even makes the light appear to originate within the crystal itself, must mean that these humble rocks possess mysterious powers. Others, while not so extreme, still find the description of a diamond as a tightly bound collection of carbon atoms repugnant, as though this knowledge somehow takes away from the beauty of the gem. In reality, a diamond's sparkle depends mostly upon human knowledge and artifice to find its expression. A rough diamond is hardly more than a dull, gray pebble, with perhaps a bit of sheen. Centuries of trial-and-error experience by diamond cutters has now been augmented by technology; a diamond to be cut is often subjected to a micrograph to determine planes along which it will most readily fracture. The various standard cuts must be carefully prepared in order that the stone show its greatest fire. It is knowledge that elicits the beauty of a diamond.

Thus the knowledge that we acquire need not preclude awe. Rather than the constricted, unchanging universe imagined by our ancestors, we now find ourselves in a dynamic and evolving universe too large for any real comprehension of its size. If some people might be distressed that humans now seem so small and insignificant, science can only respond that we are nevertheless a part of this grand cosmos, and we should feel privileged to have the ability to appreciate its true majesty. If we have been forced to abandon our anthropocentric models, in return we have gained a far grander home.

A new explanation

A new narrative

In the beginning there was neither space nor time as we know them, but a shifting foam of strings and loops, as small as anything can be. Within the foam, all of space, time, and energy mingled in a grand unification. But the foam expanded and cooled. And then there was gravity, and space and time, and a universe was created. There was a grand unified force that filled the universe with a false vacuum endowed with a negative pressure. This caused the universe to expand exceedingly rapidly against gravity. But this state was unstable, and did not last, and the true vacuum reappeared, the inflation stopped, and the grand unified force was gone forever. In its place were the strong and electroweak interactions, and enormous energy from the decay of the false vacuum. The universe continued to expand and cool, but at a much slower rate. Families of particles, matter and antimatter, rose briefly to prominence and then died out as the temperature fell below that required to sustain them. Then the electromagnetic and the weak interaction were cleaved, and later the neutrinos were likewise separated

²All true solids are, in fact, crystalline, but usually they consist of aggregates of many tiny crystals. Only occasionally does a crystal naturally grow large enough for us to appreciate its symmetry without a microscope.

from the photons. The last of the matter and antimatter annihilated, but a small remnant of matter remained. The first elements were created, reminders of the heat that had made them. And all this came to pass in three minutes, after the creation of time itself. Thereafter the universe, still hot and dense and opaque to light, continued to expand and cool. Finally the electrons joined to the nuclei, and there were atoms, and the universe became transparent. The photons that were freed at that time continue to travel even today as relics of the time when atoms were created, but their energy drops ever lower. And a billion years passed after the creation of the universe, and then the clouds of gas collapsed from their own gravity, and the stars shone and there were galaxies to light the universe. And some galaxies harbored at their centers giant black holes, consuming much gas and blazing with great brightness. And still the universe expanded. And stars created heavy elements in their cores, and then they exploded, and the heavy elements went out into the universe. New stars form still and take into themselves the heavy elements from the generations that went before them. And more billions of years passed, and one particular star formed, like many others of its kind that had already formed, and would form in the future. Around this star was a disk of gas and dust. And it happened that this star formed alone, with no companion close by to disrupt the disk, so the dust condensed, and formed planets and numerous smaller objects. And the third planet was the right size and the right distance from its star so that rain fell upon the planet and did not boil away, nor did it freeze. And this water made the planet warm, but not too warm, and was a good solvent, and many compounds formed. And some of these compounds could make copies of themselves. And these compounds made a code that could be copied and passed down to all the generations. And then there were cells, and they were living. And billions of years elapsed with only the cells upon the planet. Then some of the cells joined together and made animals which lived in the seas of the planet. And finally some cells from the water began to live upon the rocks of the land, and they joined together and made plants. And the plants made oxygen, and other creatures from the seas began to live upon the land. And many millions of years passed, and multitudes of creatures lived, of diverse kinds, each kind from another kind. And a kind of animal arose and spread throughout the planet, and this animal walked upon two feet and made tools. And it began to speak, and then it told stories of itself, and at last it told this story. But all things must come to their end, and after many billions of years the star will swell up and swallow the third planet, and all will be destroyed in the fire of the star. And we know not how the universe will end, but it may expand forever, and finally all the stars will die and the universe will end in eternal darkness and cold.

Is this a myth? If we define a myth as a narrative of explanation, it would qualify. How does this myth differ from others? For one thing, it is highly detailed. The fanciful description above is extremely condensed; the complete version of this story occupies the remainder of this book. In addition, it is not overtly anthropocentric. People play only a very limited role, even though this description was developed by humans. Nevertheless, if all you knew of this explanation was a tale such as that written above, you might have difficulty in distinguishing it from a story of ants in the tree of life. But this story differs fundamentally from the earlier myths. The most important distinction is the way in which this explanation was developed. It was based upon many centuries of observations of the universe and its contents. It draws upon the experience and thoughts of generations of thinkers, but always the most significant factor has been the accumulation and interpretation of observations. The story is held to a set of stringent constraints; it must explain known facts, and it must hold together as a coherent narrative, all the parts fitting like pieces of a grand jigsaw puzzle. How humans have arrived at this narrative, what it means, which aspects of it are more certain and which less so, and how it is to be judged, are the subject of this book. It is a lengthy story that will unfold over many chapters, but let us begin with the most fundamental basis: the establishment of criteria by which our narrative of the universe can be evaluated.

The scientific method

Scientific explanations rest upon objective data

Over the past 400 years, a new viewpoint has come to fruition, the scientific viewpoint. At first glance, this may seem to be no better than the mythology of our ancestors; it is just another belief system. However, there are significant differences between scientific and mythological explanations. In science, the ultimate judge is the empirical data, the *objective* observations. The truth, whatever it may be, is independent of humanity; but it can be known and understood, at least in approximation. The results of a set of observations, that is, of an **experiment**, must not depend upon who makes the observations. The test of any theory lies in its ability to make predictions that can be tested by further experiments. Regardless of the internal consistency of a theory, or its philosophical or aesthetic appeal, it is the data that judge the success or failure of that theory.

The development of the scientific method

The realization that the universe is knowable, at least in a practical way, developed only slowly in human thought. Although many cultures contributed to this dawning, it appeared in the first coherent way among the ancient Greeks, during the age of the philosophers some three thousand years ago. The Greeks incorporated into their system of logic the formal connection between a cause and its effect, introducing the concept, novel for the time, that a phenomenon could have a natural, consistent cause, and that cause could be identified by rational thought. The Greeks were eventually conquered by the Romans, who held the

Greek philosophy in the greatest esteem but had little real interest in furthering it themselves. After a lengthy decay, the Roman Empire itself finally collapsed in the 5th century, ushering in the Dark Ages in Europe and extinguishing nearly all the memory of the achievements of the ancients. During the Dark and Middle Ages rational thought was almost entirely absent in Europe; the knowledge gained by the Greek philosophers was preserved primarily within the Islamic world until the Crusades and increased travel and trade brought Europe into contact with other cultures once again. The Greek writings were rediscovered early in the 13th century, beginning with the works of Aristotle. Although Aristotle did much damage to scientific inquiry, both in his own time and during the late Middle Ages, the reintroduction of his texts did bring the concepts of logic and inference back into European thought, helping to pave the way for the Renaissance. During the Renaissance, and the Enlightenment that followed it, European science took shape and matured.

Science gradually became systematized. The British philosopher Sir Francis Bacon developed a procedure for scientific inquiry during the last decades of the 16th century. Bacon's methods were further refined and codified in the 19th century by a subsequent British philosopher, John Stuart Mill. Mill's Methods, as they are called, are still considered the logical foundation of science. The methods provide a formal approach to establishing inductive inferences of any kind, but science is one of their most important applications. It should be emphasized that almost all scientific hypotheses are *inductive*, not deductive. Induction is the drawing of general conclusions from an examination of particular instances, whereas deduction is the inference of particulars from general principles; the distinction between the two is often ignored in popular usage, but the difference is significant if we are to understand clearly what our observations can tell us. Since we cannot inspect every particle of matter in the universe, and our scientific laws must necessarily be based upon the data available, we must generalize from our limited experience to all the universe. Unlike deductive conclusions, which proceed from the general to the specific and can be rigorously and decisively proven, inductive hypotheses go from the specific to the general, and if the number of possibilities is too large for us to examine all of them, as is usually the case, an inductive hypothesis cannot be conclusively proven. We can, however, use deduction to test repeatedly these general hypotheses by developing specific predictions for comparison with observation.

Induction: general principles derived from specific observations

Despite the fundamental limitations of the inductive process, science has made great progress in building a consistent and comprehensible picture of the universe. The occasional failure of established hypotheses has never overturned the scientific edifice completely; instead, such failures lead to new and better knowledge of the way in which the universe works. Methodology can guide the construction of a valid (in the inductive sense) hypothesis from the known data, but cannot give a blueprint;

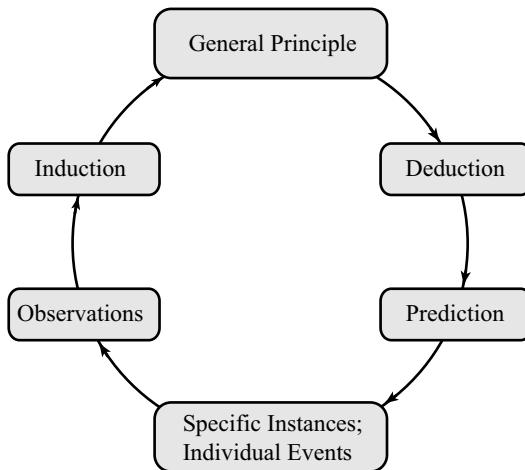


Fig. 1.3 The process of induction moves from observations of specific events to a general principle. The general principle can never be proven, since all specific instances cannot be observed. However, the principle can be tested through deduction, by which particular instances following from the general principle can be inferred.

often the great scientific hypotheses are the result of genius, hard work, or even simple luck.

Principles of the scientific method

The so-called **scientific method** is a method for testing and verifying scientific hypotheses; it proceeds, at least in principle, by several steps. First comes the gathering of **data**. We cannot build scientific explanations without careful, objective observations of the phenomenon in question; this is one of the most important distinctions between scientific and unscientific explanations. Study of the data enables the scientist to look for patterns, for similarities with other phenomena, and so forth. Once some unifying concept has been found, it may be phrased as a **hypothesis**, a working explanation for the phenomenon that can lead to further observation.

In order to be scientific, a hypothesis must have five characteristics. First, it must be **relevant**. This may seem self-evident, but it is significant. The hypothesis should be related to some observed phenomenon, not merely something invoked because the theorist happens to like it.

Second, the hypothesis must be *testable* and potentially **falsifiable**. That is, it must be possible to make observations that could support or, even better, refute the hypothesis. The importance of this characteristic cannot be overemphasized; indeed, it may be regarded as the distinguishing feature of a scientific explanation. The hypothesis that the planets are controlled by spirits was accepted for centuries, but it is not scientific because it cannot be tested; there is no observation that could disprove it. The Newtonian hypothesis, which states that planets are controlled by a force emanating from the Sun that causes them to move in specific ways, is falsifiable; if a new planet, or other orbiting body, were discovered and found not to obey the laws that Newton had discovered, his hypothesis would be disproven. On the other hand, if the new body were found to obey Newton's laws precisely, it would add evidence for the validity of the hypothesis but would not prove it.

Falsifiability unambiguously distinguishes scientific from nonscientific explanations. The philosopher of science Karl Popper put forward the

proposition that the criterion for the scientific status of a theory is the potentiality that the theory may be falsified. Pseudoscience is often based on observations, and may cite much confirming evidence, but never permits refutation. Either the contrary data are ignored, or new details are continually added to the theory in order to explain all new observations. Seen in this light, the scientific status of a theory is granted not so much by its explanations, but by its prohibitions: the theory says what cannot happen, and if those things are observed, then the theory is wrong.

A scientific hypothesis must also be **consistent** with previous established hypotheses. If a known hypothesis explains a phenomenon well, and has passed many experimental tests, we would be ill advised to abandon it merely because a newer and shinier explanation might appear. This principle is often little appreciated by the public, or by pseudoscientists who cite Einstein or Galileo as iconoclasts who refuted established science. In fact, Einstein's theory of relativity would not have been accepted had it not been consistent. Newton's laws of motion were well established even during his own lifetime as a very good explanation of mechanics. Over the next three centuries, they were verified time and again. Yet there remained one nagging problem, which Albert Einstein set out to solve. In doing so, he was forced to give up notions about the universe that had been cherished for centuries, but which were not essential to understanding the Newtonian observations. The special theory of relativity revolutionized our conceptual view of space and time and showed itself to be a more *complete* theory of motion, in that, unlike Newton's laws, it was applicable at all speeds, and it made electromagnetics consistent with mechanics. Nevertheless, the special theory of relativity is fully compatible with Newtonian mechanics, and can be shown to reduce to the Newtonian theory for all material motions at speeds well below the speed of light. This is precisely the regime in which Newton's laws were known to be valid to within the accuracy of the data available. Einstein did not *refute* Newton, but rather he modified and *extended* the laws of mechanics into previously unexamined and untested domains.

The criterion of consistency is important, but not absolute. It *is* possible for an old theory that is well accepted to be simply wrong, and a new one replaces it completely; but such incorrect theories survive only in the absence of data. A good example of this is the caloric theory of heat. For many years, heat was believed to be some sort of invisible fluid that flowed from a hotter to a colder body. The caloric theory was able to explain many common properties of heat reasonably well. It was not until more careful measurements were made and better data were collected, beginning with Count Rumford's observations of cannon-boring at the end of the 18th century, that the theory was called into doubt. In 1799, Sir Humphrey Davy conducted a **crucial experiment**, one which has the power to decide between two competing theories on the basis of a single incompatible prediction. Unfortunately his experimental design was somewhat lacking, and his results were not convincing. But the

Science versus pseudoscience

Scientific knowledge is cumulative

A crucial experiment is designed to distinguish between alternative hypotheses

way was shown, and within fifty years several scientists, most especially Sir James Joule, developed the kinetic theory of heat, which is accepted today. The new theory was incompatible with the old, and the caloric theory was discarded.

A fourth criterion for a scientific hypothesis is **simplicity**. This is a somewhat subjective criterion, to be sure, but it has guided the development of many theories. All other things being equal, the simpler explanation is favored, an assertion often known as **Occam's Razor** for the medieval English philosopher William of Occam (or Ockham) who asserted that “entities must not be needlessly multiplied.” A good theory does not require a special rule for each observation.

The fifth important criterion for judging a scientific explanation is its **predictive power**. Predictive power is not quite the same thing as falsifiability, although the two are interrelated. Predictive power refers to the ability of the hypothesis to predict new, previously unobserved phenomena. Similar to this, and part of the same criterion, is the **explanatory power** of the hypothesis, which is a quantification of the number of facts the hypothesis can encompass and explain. Given two otherwise similar hypotheses, the one with greater explanatory power is generally preferred. Predictive power is even better, for then the hypothesis can be bolstered if the new phenomenon is observed, or discredited or even disproved if the phenomenon is not observed, or is observed but behaves contrary to the prediction of the hypothesis.

In order to be accepted, any new hypothesis must represent an improvement. It must explain more facts, or provide a better explanation of the existing knowledge, than does the older theory. Although great theories are often advanced by individuals, science as a whole is a social activity. It is not the brilliance or authority of one person that forces the acceptance of a hypothesis. Although hypotheses, like clothing, may come into fashion or fall from favor for all-too-familiar human reasons, such as dominance by one powerful individual or a scientific fad, it is inevitable that over time, only those explanations that can win the acceptance of the scientific community prevail. And by the communal nature of science, such hypotheses must fit in with the overall picture in order to win any such contest.

If a hypothesis becomes especially well established and survives many tests that could have refuted it, it may be elevated to the status of a **theory**. A theory, in strict scientific usage, is a hypothesis that is sufficiently accepted and which shows enough explanatory power to be strongly confirmed by experiment. It is *not* a conjecture, as the word theory often connotes in popular usage, where it has little more import than an opinion. Occasionally an especially well confirmed theory is called a **law**, but this usage has diminished considerably in the past century. The terminology is by no means consistent and in any case, most scientific explanations, being inductive, are necessarily hypotheses with lesser or greater degrees of verification. However, in no case is a scientific hypothesis or theory a mere guess. It is always founded upon a careful methodology for correct inductive inference, and it is judged by

The power of scientific theory

A scientific theory represents our best statement of fact

the criteria we have described. Ultimately, the data decide. No matter how beautiful the theory, its success or failure is determined by how well it explains our observations, both those already known and those that will come from experiments yet to be performed.

The progressive nature of science should also make clear that it does not seek a revealed or absolute truth, but instead looks for models of reality. A **model**, in this context, refers to the coherent description established to explain a phenomenon. It is more or less equivalent to a theory; that is, it is an established hypothesis or set of interrelated hypotheses. For example, the big bang model of the universe is a mathematical construction that provides illumination and interpretation for the data we collect. This does not mean that a model is a fiction that has nothing to do with reality; on the contrary, in modern science, a model represents the best description of the phenomenon that we can devise, and insofar as it succeeds at reproducing the observations, it surely must touch some facet of reality. It does mean that a model never claims to *be* reality. If better data invalidate part or all of our model, we must replace it appropriately. The failure of a model does not represent a failure of science; science fails only when we cling stubbornly to a model that has clearly ceased to be the best possible.

A model must never be confused with the entity it represents. No matter how good a cosmology we may eventually develop, it is still a product of the human mind, yet we would not claim that the universe is a human construction. Humans have strong intellectual gifts, especially with our unique ability to consult with one another, but our brains are still finite; it may be that some aspects of reality are beyond our grasp. Even if physicists develop an ultimate theory that explains all that can be known about elementary particles, this will tell us little about how consciousness arises, or about a host of other complex problems. Reality may be a fleet runner we can never overtake, but which we can approach ever closer.

Despite the grandeur of its subject, modern cosmology is a science and obeys the rules of the scientific method. Cosmologists formulate hypotheses and appeal to data to test them. Cosmology is primarily an observational science, as opposed to an experimental science. We cannot arrange to perform our own experiments on the universe, controlling them as we like, but must be content to observe what we happen to see. Cosmologists attempt to tie those disparate observations together with physical theory to create the best cosmological hypotheses possible. These new hypotheses may then suggest new observations, as a good scientific hypothesis should do, and from those observations we may strengthen or discredit the explanation. Thus we humans make cosmological progress, despite our confinement to the immediate vicinity of a small planet orbiting a modest star in a run-of-the-mill galaxy. Yet even from our restricted vantage point we shall find a universe more wondrous than our ancestors, with their capricious gods and their pre-occupations with geometrical or mystical perfection, could ever have dreamed.

All models have limits of validity

Cosmology is a science

Chapter Summary

Humanity has long sought an understanding of the cosmos. Mythology, humanity's first attempt to grapple with cosmological questions, consists of narrative tales that describe the universe in understandable terms. Cosmology, particularly as expressed by a mythology, can influence a culture's or an individual's actions. The big bang appears to the casual observer as just another myth, albeit without some of the more obvious anthropocentric characteristics. The difference, however, is that modern cosmology is based upon the scientific method. The scientific method has very specific rules. It is based on objective data, observations that are independent of the observer. Once sufficient data are collected, a hypothesis is framed to explain and unify them. To be regarded as scientific, the hypothesis must meet at minimum five characteristics: it

must be relevant, testable, consistent, simple, and possess explanatory power. Of these, the property of testability particularly defines the scientific method. A hypothesis that does not contain the potential to be falsified is not scientific. Once a hypothesis has met success at explaining data and has proven itself useful in predicting new phenomena, it is generally called a theory. Some particularly well established theories, especially those pertaining to a limited phenomenon or forming the foundation for a broader theory, are called laws. Hence we refer to the law of gravity, even though scientific laws are subject to modification as our understanding improves. A model, which is more or less equivalent to a theory or a set of interrelated theories, can be constructed to produce the best explanation possible of a particular phenomenon.

Key Term Definitions

cosmology The study of the origin, evolution, and behavior of the universe as a whole.

universe That which contains and subsumes all the laws of nature, and everything subject to those laws; the sum of all that exists physically, including matter, energy, physical laws, space, and time.

astronomy The study of the contents of the universe beyond the Earth.

anthropomorphism The projection of human attributes onto nonhuman entities such as animals, the planets, or the universe as a whole.

myth A narrative intended to explain or justify the beliefs of a people. The term usually suggests a lack of historical and factual basis.

anthropocentrism The belief that humans are central to the universe.

experiment A controlled trial for the purpose of collecting data about a specific phenomenon.

scientific method An investigative approach in which data are gathered, a hypothesis is formulated to explain the data, and further experiments are performed to test the hypothesis.

data The outcome of a set of measurements from which inferences may be drawn, theories constructed, etc.

hypothesis A proposed explanation for an observed phenomenon. In science, a valid hypothesis must be based upon data and must be subject to testing.

relevant Of a scientific hypothesis: directly related to the phenomenon it seeks to explain.

falsifiable Of a scientific hypothesis: leading to the possibility of performing an experiment that would disprove, or falsify, the hypothesis.

consistent Of a scientific theory: containing and extending an earlier well-supported theory, e.g. general relativity is consistent with Newtonian gravity.

crucial experiment An experiment that has the power to decide between two competing theories.

simplicity The property of a scientific hypothesis that its proposed explanation must not be unnecessarily complicated.

Occam's Razor The principle that when all other things are equal, the simplest explanation is preferred.

predictive power The ability of a hypothesis or model to predict unobserved effects. This provides an important means of testing a hypothesis.

explanatory power The ability of a scientific hypothesis to account for known data.

theory In scientific usage, a hypothesis or related group of hypotheses that have become well established.

law In scientific usage, a theory that has become particularly well confirmed and well established.

model A hypothesis or group of related hypotheses that describes and clarifies a natural phenomenon, entity, etc.

Review Questions

- (1.1) For at least one myth, either one from the text or one of your own choosing, identify the major theme(s) and explain how the myth fitted the social and political circumstances of the people who developed it.
- (1.2) Give an example of how one or more cosmological assumptions have influenced the behavior of modern political leaders in an industrialized nation.
- (1.3) Find your horoscope for one particular day in a newspaper. Keep track of your activities for the day, observing any occurrences that could appear to be fulfillments of the horoscope. Did anything happen that was explicitly contrary to the predictions?
- (1.4) Repeat the activity in Question 3, but for a horoscope that is *not* yours, and is chosen randomly from the horoscopes separated from yours by at least two houses. (Recall that the ordering of the houses is circular; Aquarius follows Capricorn.) Ideally, this and the preceding exercise should be done with the help of a friend, so that you do not know whether the horoscope you are given corresponds to your birthdate or was randomly selected.
- (1.5) Describe at least two examples of anthropocentric beliefs that are still widespread.
- (1.6) What is an experiment, and what is its role in science?
- (1.7) Explain the distinction between inductive and deductive reasoning.
- (1.8) Describe the five major criteria for evaluating scientific hypotheses. Which are most important? Why?
- (1.9) Define the word *theory* as it is used in science. How does this usage differ from a common everyday meaning of the word?
- (1.10) Choose an example of a pseudoscientific theory and explain how it fails to be falsifiable.
- (1.11) What is the ultimate arbiter of truth in science? How does this distinguish science from other systems?

This page intentionally left blank

Cosmology Becomes a Science

2

I demonstrate by means of philosophy that the earth is round, and is inhabited on all sides; that it is insignificantly small, and is borne through the stars.

Johannes Kepler, *Astronomia nova*

For thousands of years, the universe that occupied human minds was small, limited by human senses and abilities. The world seemed to end at the horizon, and few traveled far from the towns of their births. The heavens were the realm of gods, beyond the understanding of mortals. From such a narrow perspective, it is not surprising that the universe appeared to be dominated by the Earth. The stars were held to be eternally fixed in their positions on the celestial sphere. The “wanderers,” or planets, known from ancient times and by nearly all cultures as entities distinct from the fixed stars, were thought to be under the control of, if not literally the embodiment of, gods or spirits. The Earth was apparently motionless, while the sky and planets, including the Sun and Moon, revolved around it. But if the Earth is still and everything else moves, is it not perfectly reasonable to conclude that the Earth is the center of the universe?

This was the dominant cosmology in Europe from ancient, perhaps prehistoric, times until the close of the Middle Ages. Then, over an astonishingly brief span of less than two centuries, the prevailing worldview changed dramatically and irrevocably, bringing about what is often called the *scientific revolution*. Over the past 300 years, further elucidation of the new cosmology has continued, bringing us to our modern models. The new universe that has emerged might seem as strange to Isaac Newton as would his to ancient philosophers.

Greek cosmology

More than 2,000 years ago, Greek philosophers developed a sophisticated system of rational thought, establishing the basic rules of deductive logic that are still followed today. Some of the early philosophers were also scientists, performing feats of astronomy that, in light of their extremely limited ability to make quantitative observations, seem impressive even

Key Terms:

- geocentric
- mechanics
- force
- inertia
- heliocentric
- parallax
- retrograde motion
- Copernican revolution
- Copernican principle
- ellipse
- Kepler's laws
- thought experiment

Prescientific cosmologies were limited in scope

The roots of modern science can be traced to Greek culture

now. When Greek culture was temporarily forgotten, European thought degenerated into the superstition and fear of that dismal period known as the Dark Ages. The rediscovery of Greek culture, as well as the discovery of the achievements of other Mediterranean cultures, led ultimately to the Renaissance.

Today we acknowledge that Western science has its roots in Greece. The Greeks did not invent their system from nothing, but were influenced by neighboring peoples; however, it was they who were chiefly responsible for establishing the basic principles of scientific inquiry. Among their accomplishments was the identification of cause and effect. This may seem obvious to us now, but it was an important conceptual advance and an essential prerequisite for scientific thought. The Greeks realized that it was possible to observe a natural phenomenon and to seek an explanation for the observation. It was even possible to understand nature in precise mathematical terms, which meant geometry to the ancient Greeks. To move from an understanding of Earthly phenomena to a grasp of the universe is then merely a matter of scale. If we can measure the size and shape of the Earth, we can do the same for the heavens. With the concepts of cause and effect in place, the world is no longer random and capricious; instead, it is ordered and predictable.

The predominant feature of the mainstream Greek cosmology was the centrality of an unmoving Earth. As remarkable as it may seem, the spherical shape and the size of the Earth were well known to the Greeks. Despite the restricted ability of the ancients to travel, they were aware that the view of the constellations, at the same time of year, changes as one moves north or south. More evidence was found in the fact that ships with tall masts disappear as they move away from the coast, but not in a proportional manner; first the hull drops from view, then only later the mast. This would not happen on a flat plane, as geometers could appreciate; thus they concluded that the surface of the Earth must be curved. Furthermore, the Greeks had also deduced the cause of lunar eclipses, and realized that the shadow of the Earth on the Moon was curved. Once the shape of the Earth was determined, it became possible to ascertain its size. The Greek geometer Eratosthenes (circa 3rd century BCE) computed the diameter of the Earth by measuring the altitude of the Sun in the sky at two different locations on the Earth at noon on the summer solstice. With the reasonable assumption that the Sun's rays were parallel, he was able to use these measurements to obtain a result that historians believe to be quite close to the correct figure. To surround this spherical Earth, the ancient Greeks supposed that the sky too was a physical sphere; they believed that it hung overhead, relatively close to the Earth.

An important factor in establishing the Greek cosmos was the conclusion that the stars moved, whereas the Earth did not. Motion had long been recognized in the heavens; the patterns of stars changed with the seasons, while the planets, including the Sun and the Moon, moved among the stars. But for observers confined to its surface, all available evidence indicates that the Earth itself does not move. If the Earth,

A cosmology centered on a spherical Earth

Does the Earth rotate?

rather than the celestial sphere, were turning, then near the equator a point on the surface would have to be moving at the incredible speed of nearly 1600 kilometers per hour. Surely such speeds would be perceptible! Would not such a great motion generate winds with enormous velocities? Moreover, how could someone jumping from the surface of the Earth land in the same spot from which he leaped? When one is standing upright, there is no sensation of motion. A dropped object falls straight down. The concept of a moving Earth also seems to conflict with the observation that moving objects tend to come to rest, and to remain at rest unless impelled. How could the Earth sustain movement, when all other Earthly motions rapidly come to a halt? But if the Earth is stationary and everything else in the universe revolves around it, then obviously the Earth must be the center of the universe. To the ancients, arguments such as these established unambiguously the motionlessness and centrality of the Earth. A cosmology that places the Earth at the physical center of the universe is said to be **geocentric**. The geocentric model of the universe fit perfectly with the anthropocentric attitudes that dominated, and in many ways still dominate, most human thought. Indeed, the geocentric model remained the mainstay of cosmology until scarcely 400 years ago.

Once the basic structure of the universe was decided, the next task was to describe and explain the heavenly motions. The model of the celestial motions must do more than provide a general description; it should make detailed predictions that would be as accurate as the observations. The difficulty, as any casual student of the heavens rapidly comes to appreciate, is that the motions in the sky are intricate. The prejudices of the ancient Greek geometers for certain figures further complicated their construction of a model. The scientific process has always required an interaction between ideas (theory) and observations (data and experiment). Today we regard accurate observations as supreme; theory must give way if need be. The Greeks felt the opposite to be true. Theory, which sprang from pure rational thought, was considered to be superior to observations, which were sullied by the unreliability of human experience and senses. Thus the early Greek scientists felt no qualms about forcing the universe to conform to their philosophical ideals.

The first systematic cosmology we shall consider was developed by the philosopher Plato and his student Eudoxus in the 3rd century BCE. Eudoxus set out to create a system that adequately agreed with real observations, while preserving accepted ideas about motion and geometry. The resulting cosmological picture, later considerably refined by Aristotle and others, was based upon the sphere, which, according to Greek philosophy, was the most perfect of solid geometric forms; correspondingly, the circle was the most perfect curve. The sphere has the appealing property that it encloses the largest possible volume for a given surface area, an aesthetic much appreciated by the Greeks. Justification for the spherical universe was also found in the recognition that the Earth itself was a sphere; surely this shape was no accident, but

Spheres and circles were considered the perfect forms

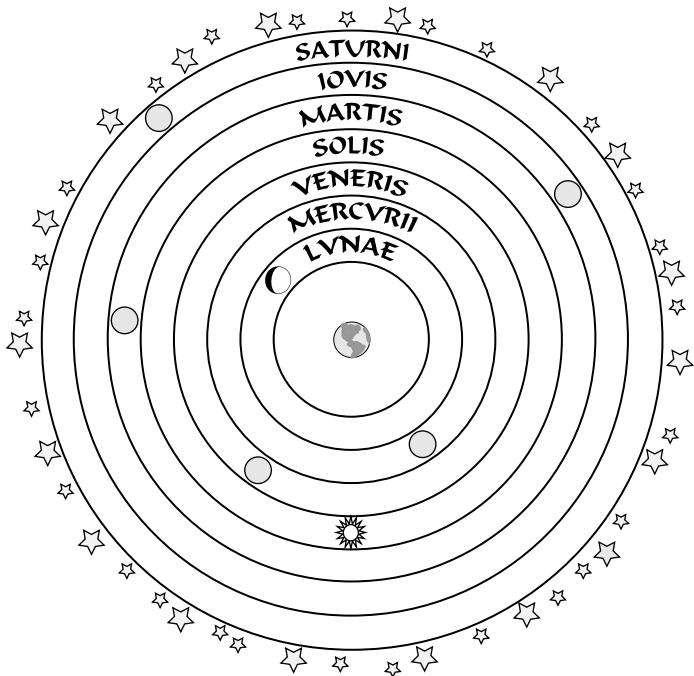


Fig. 2.1 The simplest geocentric model of the cosmos. The universe is finite, with Earth at the center, surrounded by the spheres of the Sun, the Moon, and the planets. The sphere of the stars lies at the outer edge. In reality, the motions of the Sun, Moon, and planets cannot be adequately described by a single sphere for each. The models of Eudoxus and his successors postulated multiple spheres for each of these bodies.

reflected the geometrical design of the universe. We should not criticize the Greeks for relying so heavily upon their notions of symmetry, for even modern physicists profess a great appreciation for symmetry in their theories. In modern science, however, symmetry is a guide and not an arbiter, and the symmetries invoked are often quite subtle.

The obvious approach to constructing a cosmos based upon spheres, centered on a spherical Earth, requires separate spheres for the Sun, the Moon, and each of the planets; only one sphere is required for all of the fixed stars, which move as a unit. Unfortunately, the motions of everything but the fixed stars is more complex than can be accommodated within such a straightforward model. Even at the time of Eudoxus, observations were adequate to rule out a simple circular motion of the planets about a stationary Earth. Multiple spheres are needed to account for the various observed motions of even one celestial body. For example, the Sun exhibits its familiar daily motion through the heavens, for which a single sphere, rotating on a 24-hour schedule, can account; but the Sun also has a longer seasonal motion as it moves north and south of the Equator. The seasons, therefore, require a second sphere. The more complex motions of the Moon and the planets required even more spheres. In the end, Eudoxus was obliged to introduce 27 different celestial spheres, each with a different rate of rotation and orientation of its axis. The result was less geometrically beautiful than it was practical; it fitted the observations of the day reasonably well. Eudoxus' model set the pattern for future refinements.

Aristotle

The most famous of Plato's students was not Eudoxus, nor was Eudoxus the only one who pondered cosmology. By far the most influential of the ancient Greek philosophers was another pupil of Plato, Aristotle (384–322 BCE). Aristotle wrote widely and voluminously on nearly every conceivable philosophical issue of his day. Much of what we know of Greek cosmology comes from his writings, and from the work of later members of his school, who edited and revised his texts. Although he made many original contributions to a number of fields, and was one of the first to develop a theory of biology of any kind, most of the elements of Aristotelian cosmology are common to other cosmologies of the era. The writings of Aristotle became particularly influential because he justified his cosmology on rational grounds.

Aristotle developed his model within a general physical theory. This remains the basic approach in modern scientific cosmology. If the universe has certain properties and the objects within it behave in certain ways, then there must be principles behind these behaviors: specifically, natural laws. The universe embodies these laws, and they can be discerned by humans if enough observations are made. It is sometimes claimed that Aristotle developed his theories by thought alone, without regard to observations, but this is not true. He made observations, to the best of his ability, then attempted to reason from those observations. In this respect, his work represented a break with the earlier Platonic school of thought, which held that truth lay only with ideas. To Plato, observations were misleading because the physical world was at best a pale manifestation of the truth; only pure geometry could claim to represent the ultimate reality.

An important element of Aristotle's cosmology was his theory of motion; today we call this branch of physics **mechanics**. Motion, taken mostly for granted before his time, presented many questions to Aristotle. Why do objects on the Earth have the tendency to move as they do with respect to the Earth? Why do objects fall when dropped, and why do stones sink in water, while bubbles rise? It must be, thought Aristotle, because it is in their fundamental natures to move so. In many ancient theories, all objects are composed of the four basic elements of earth, water, air, and fire. In the Aristotelian view, each of these elements was believed to move differently: earth toward the center, fire away from the center (flames rise), while water and air occupy the space between. Air bubbles up through water, but rocks sink. Consequently, objects of different compositions fall at different rates. An object containing a higher proportion of the lighter elements air and/or fire would fall slowly, whereas an object consisting mostly of earth would fall quickly. The conclusion that various bodies fall at different rates was consistent with casual observation. The composition or nature of the object thus determined its mechanical behavior. All things sought to move to their natural place in the cosmos. Because of this, all motion must be with respect to the basic structure of the cosmos; specifically, Aristotle

Aristotle's cosmology was based on a physical theory

Aristotle's natural motion defined

proposed that Earthly objects move in a straight line, that is, *linearly*, with respect to the center of the universe. The Earth is a sphere, so objects falling straight down are actually moving toward the center of the sphere, as Aristotle realized. Therefore, the center of the cosmos must lie at the center of the Earth. This argument provided a rationalization for the geocentric cosmology based not only upon celestial motions but upon a physical theory.

Deviation from natural motion requires a force

Natural motion is thus defined within the Aristotelian model. But what of other motions? Aristotle's law of motion incorporated the idea that **force** causes a deviation from natural motion, a significant advance in understanding. The concept of force is intuitive; it is a push or a pull, an action by one thing on another. This concept still remains a fundamental part of modern mechanics, although in a much more quantitative form. Aristotle observed that a force is required to set a stationary object into motion, and that Earthly motions tend to die out shortly after they are initiated. For example, a rock thrown, no matter how energetically, soon falls to the ground and stops. Aristotle proposed that a force is required to make an object move in any manner different from its natural motion. A horse must continually pull on a cart to move it. Similarly, an arrow shot from a bow must experience a sustained force during its flight, or so Aristotle thought. Aristotle believed that objects in flight, such as an arrow, were somehow pushed along in their paths by the air, with a kind of highly localized wind.

The law of inertia

Although ground-breaking, Aristotle's law of motion was erroneous. Viewed with the hindsight provided by modern physics, we can see where he went astray. Aristotle's difficulties arose because he only partially grasped the concept we now call **inertia**, the tendency of a body to resist changes in its state. He realized that a body at rest will remain at rest unless impelled by a force; but he missed the other, equally important, part of the law of inertia: a body in motion in a straight line will remain in that state unless a force is exerted. From modern physics, we know that a force is required to produce a *change* in a state of motion. To Aristotle, continuous motion required the continual application of force; he could not conceive of the possibility that an Earthly object might travel forever on its own.

Circular motion in the heavens

But what of the heavenly motions? In contrast to Earthly motions, celestial motions *do* continue indefinitely. The motions of Aristotelian heavenly objects cannot follow straight lines, since straight lines would end at the edge of the universe, and thus all such motion would ultimately be finite. Hence there must be two separate types of natural motion: straight-line (linear) limited motion in the Earthly realm, and continuous circular motion in the heavens. This was one of Aristotle's most influential axioms: the *primacy of circular motion* in the heavens. It made a certain geometric sense: lines are of finite length, whereas a circle closes back upon itself and has neither a beginning nor an end. Because the heavenly bodies had a different natural motion, circular and eternal, they could not be composed of earthly materials, which could move only linearly toward their appropriate place in the cosmos as de-

terminated by their composition. Instead, celestial objects were composed of *ether*, a fifth element. Since they were already in their proper place with respect to the center of the cosmos, they moved in perfect circles. Heavenly bodies would thus continue to move indefinitely without the action of any force. Aristotle argued that the ethereal heavens were eternal and unalterable, perfect in their structure and unchanging. The Earthly world below changed, but the heavens did not. Any apparent change in the heavens must therefore be linked to the Earth. Aristotle argued on these grounds that meteors and comets were manifestations in the upper atmosphere of the Earth.

According to Aristotle, the Earth was surrounded by nested, crystalline (transparent) spheres of the heavens, to which were attached the celestial bodies. Whereas Eudoxus apparently thought of the spheres as mathematical entities only, useful for description but not to be taken literally, Aristotle gave them physical reality and a composition. These spheres rotated around the Earth, carrying the heavenly bodies with them. This spherical universe of Aristotle was consistent with the physical and philosophical reasoning of his time, but the final model lacked much of the aesthetic quality that had originally motivated the Greek philosophers. Alas, the geometrical beauty of spheres and perfect circular motion encountered the obstacle that plagues all theories: better observations. In order to meet the challenge of the observations of the day, Aristotle was obliged to postulate 55 separate spheres to account for the motions of a far smaller number of bodies.

How large was Aristotle's grand construct? The size of the universe was limited by its fundamental geocentric property. The heavens were moving, not the Earth. Consequently, the universe must be finite, for an infinite universe rotating around a center would necessarily travel an infinite distance in a finite interval of time. In Aristotle's cosmology the distance to the stars is very small (by modern standards) in order to prevent them from moving at unreasonable speeds. The entire Aristotelian universe would fit comfortably into a region smaller than that defined by the Earth's orbital radius around the Sun. This finite universe had an edge, but it could never be reached because any motion toward the edge would shift from linear to circular as the traveler approached the heavenly realm. Even though space was assumed to have an edge, Aristotle apparently could not imagine an edge to time, so he took the point of view that time must be infinite, without beginning or end. The Greeks were aware that recorded history did not stretch back to infinity, and that change occurred. This fitted into the philosophy of the Earth as imperfect, made of four base elements, while the heavens were composed of eternal, perfect matter. The Earth changed, while the heavens did not. Conversely, the Earth did not move, while the heavens did. Few philosophers of the time, or even much later, seemed inclined to question why the center of a perfect universe would be located on an imperfect, woeful planet.

Aristotle's cosmology was very much a product of its time and of its author. While it cleaved tightly to the ancient view of the Earth and

A framework for a spherical, geocentric cosmology

sky in its insistence upon an unmoving, central Earth and a perfect heaven, it still contained important, original contributions. As we have suggested, the supreme accomplishment of Aristotelian cosmology was the argument that the universe could be described in terms of natural laws that could be inferred through rational thought. Aristotle founded the science of mechanics, and developed the concept of force into something that was at least vaguely systematic. In Aristotle's cosmology, the structure of the universe is inextricably linked to physics and to the definition of natural motion. Remarkably, this is true for modern cosmology as well. Cosmology cannot exist as a science without physics; the general structure of physical theory affects the underlying cosmology and *vice versa*. As humanity's understanding of physics improved, first from Newton and later from Einstein, the universe changed as well.

Flaws in Aristotle's theories

Unfortunately, Aristotle's work was also fundamentally flawed. The Aristotelian laws of motion did not include the correct concepts of natural motion or inertia. Also influential, but quite wrong, was his separation of the universe into Earthly and celestial realms, governed by separate laws and composed of separate elements. These misconceptions, especially the demand that celestial motion be circular, would confuse and confound physics, astronomy, and cosmology for seventeen centuries. However, we cannot blame Aristotle too much for developing a physics that was largely incorrect. The fault in his method was that his observations were often misleading. For example, he did not understand phenomena such as friction or air resistance, nor was he able to recognize that if all objects on a surface are moving with that surface, the motion will be difficult, perhaps even impossible, to detect locally. Perhaps most importantly, he did not recognize his own limitations, both as an observer and a theorist. Aristotelian physics matched the intuitive beliefs of most people and suited their philosophical leanings as well. Consequently, the geocentric theory was retained and enshrined, eventually reaching the point of religious dogma during the Middle Ages.

Not only the Aristotelian cosmology was venerated during the Middle Ages; the corresponding Aristotelian physics of motion was further elaborated into the *impetus* theory. In this view, objects moving on the Earth are propelled by an impetus, a vaguely defined, traveling, generalized force. For example, the impetus theory holds that a rock shot from a catapult is endowed with some amount of impetus that continues to propel the rock forward. The rock falls back to Earth when it has consumed all the impetus provided by the catapult. Similarly, in the case of an arrow shot from a bow, the medieval theory took the arrow to be pushed not by any vortex of air, as Aristotle had believed, but by impetus imparted to it by the bow. In this picture, air resistance is a factor acting on bodies to exhaust their impetus; the more massive the object, the faster the resistance dissipates its impetus. Impetus was also hypothesized to follow the form of the original motion; if a ball was whirled in a circle and then released, it would carry circular impetus with it and thus would continue to execute curved motion.

Modern researchers into science education have found that many people hold an intuitive view of the world that is very similar to that at which Aristotle arrived, and that contains many elements of the later impetus theory. Most people consistently misinterpret observations of motion; a notion of something like impetus still governs the way in which many of us think about motion. For example, when shown a ball traveling along a spiral track toward an exit and asked to describe the ball's motion after it leaves the track, many people believe that the ball will continue in a circular path. As we shall see when we study Newton's laws, this is incorrect. When we observe the flight of a ball or arrow, effects such as air resistance, or aerodynamic lift provided by a ball's spin, alter the trajectory in complicated ways. It is very difficult to derive the true laws of motion from our observations of such everyday occurrences.

Impetus theory seems consistent with intuition

Perhaps even more remarkable has been the survival of the Aristotelian distinction between the Earthly and the celestial realm. Despite the great gains in understanding over the past few centuries, this viewpoint lingered even into the modern era. Newton demonstrated in the 17th century that celestial motion was governed by the same laws as Earthly motion, yet space remained a mysterious realm. Prior to any manned spaceflights, exaggerated scientific and medical concerns about grave dangers were voiced. For example, fears that astronauts might go insane merely from being exposed to outer space could well have been a relic of the Aristotelian cosmology. Prominent scientists expressed opinions that a Moon landing would be extraordinarily dangerous because of deep seas of dust, or Moon germs, or highly reactive compounds in the lunar soil that would burst into flame when first exposed to oxygen. These concerns were put to rest in a most decisive way: humans visited the Moon, and in just a few years transformed it from the exotic to the mundane. Television transmissions of astronauts bouncing about on the Moon showed it to be a real physical object, made of rock, covered with fine dust, interesting but also familiar. Going beyond the Earth's immediate vicinity, photographs sent back by the *Viking* landers from the surface of Mars resembled scenes of terrestrial deserts. By now spacecraft, such as *Voyager* I and II, have visited all of the worlds of our solar system except Pluto. We have found that each planet and moon is unique, with its own history and geology, yet each is a physical world, obeying the same natural laws as does the Earth.

The celestial and the Earthly realms are not distinct

Heliocentrism ahead of its time

Aristotle was by far the most influential of the ancient Greek thinkers, especially among later Europeans. Nevertheless, his theories were not the only ones developed by Greek scholars. Ancient scientists who belonged to competing philosophical schools, especially the Pythagoreans, were making remarkable progress with the very limited tools, both mathematical and observational, that were available to them. One of the most outstanding of these Greek scholars was Aristarchus of Samos (ca. 310–

Celestial calculations from naked eye observations

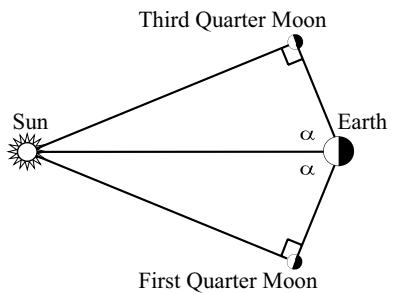


Fig. 2.2 Method proposed by Aristarchus to measure the distance from the Earth to the Sun. The angle α can be determined from the ratio of the time interval from third to first quarter to the interval from the first to the third quarter. At the quarter phase, the Earth, Moon, and Sun form a right angle. Simple geometry yields the Earth-Sun distance.

The first known heliocentric cosmology

230 BCE). Aristarchus came close to the modern description of the solar system, a millennium and a half before Copernicus. Aristarchus arrived at his model when he set out to calculate the relative sizes of the Earth, the Sun, and the Moon, using geometry and eclipse data. The relative sizes of Earth and Moon can be determined by comparing the shadow of the Earth with the angular size of the Moon during a total lunar eclipse. From these data, Aristarchus was able to conclude that the Moon had approximately one fourth the diameter of the Earth, very close to the correct ratio. He also obtained a very accurate value for the distance from the Earth to the Moon.

Obtaining the distance from the Earth to the Sun is more difficult; in fact, this measurement was carried out to good accuracy for the first time only in 1769, after dramatic improvements in knowledge and technology made it possible to exploit for triangulation the rare passage of Venus directly across the face of the Sun. Aristarchus instead used a method that was extremely clever, although difficult to make work in practice: he attempted to triangulate on the Sun by using the phases of the Moon. When the Moon is in its first or third quarter, that is, half its surface is illuminated, the angle defined by the lines from the Earth to the Moon, and the Moon to the Sun, is a right angle. The other angle required for the triangulation is proportional to the ratio of the time elapsed between first and third quarter, and third and first quarter. The closer the Sun is to the Earth, the shorter is the time elapsed between the third and first quarters of the Moon, in comparison to the corresponding interval between the first and third quarters. Unfortunately, Aristarchus could not have carried out an accurate determination with this technique, as he had neither a precise method of detecting, by naked-eye observation, the exact moment at which the Moon is exactly half illuminated, nor did he possess accurate clocks to measure the time intervals required. Even with these obstacles, Aristarchus obtained a distance to the Sun of 19 times the distance to the Moon. This number is much too small, by about another factor of 20 (the correct result is that the Sun is 390 times as far as the Moon), but the geometry was sound, and Aristarchus was led to an incredible conclusion. He knew that the Sun and Moon had the same apparent size in the sky, from the remarkable fact that the Moon precisely covers the Sun during a solar eclipse. By his measurement, the Sun was roughly 20 times as distant as the Moon; therefore it must be 20 times the diameter. Since the Moon was one quarter the diameter of the Earth, the Sun must be much larger than the Earth. This led Aristarchus to propose the first **heliocentric** cosmology, in which the Sun, not the Earth, was the center of the universe.

Aristarchus' heliocentric model was never accepted by his contemporaries, who raised what they considered to be sound objections against it. First, it required that the Earth move, in violation of both sensory evidence and the prevailing physics of the day. Second, a moving Earth has definite observable consequences. Since the Greeks believed that the stars were located on a relatively nearby celestial sphere, the Earth's orbital motion should bring different regions of that sphere no-

ticeably closer at certain times of year; no such stellar brightening was seen. Moreover, over the course of a year the stellar positions should shift as the stars are viewed first from one side of the Earth's orbit and then from the other. This phenomenon, known as **parallax**, had never been observed in Aristarchus' time. The only way in which the absence of parallax could be explained within the context of the heliocentric models was to demand that the stars be at enormous distances from the Earth (which, of course, they are). Aristarchus' cosmos was, by the standards of his day, fantastically huge, with a radius comparable to the distance we now call a lightyear. We now know that this is barely a quarter of the distance to the nearest star, but at the time this immense size could not be accepted by most people. Aristarchus, who was probably one of the most brilliant of the ancient scientists, was too far ahead of his time. His theory probably also did not win favor because people were not yet ready to accept that their Earth was not the center of the universe and the sole preoccupation of its gods.

Ptolemy

The work of the Greek astrologer and geographer Claudius Ptolemaus, called Ptolemy (ca. AD 100–170), brought the Aristotelian system to its pinnacle. Ptolemy worked in an observatory near Alexandria, the great seat of learning of ancient Egypt. His principal work, the result of his years of study, is generally known by its Arabic name *Almagest* (The Great System). This opus brought together all the refinements of Aristotelian cosmology to describe better the observed motions of celestial objects. Ptolemy was not only a theorist, but spent time charting the movements of the stars and refining his system to fit his observations. Moreover, by his time a long history of observations had accumulated, exposing the inadequacies of earlier models. Slight inaccuracies in predicting conjunctions of Jupiter and Saturn might not be noticed over the course of a few years, but over a few hundred years these errors would become substantial. Of course, the Greeks had known long before Ptolemy's time that the simplest possible geocentric system, in which each planet describes a circular orbit around the Earth, could not fit the data. Eudoxus' 27 spheres and Aristotle's 55 were the consequence of this celestial intricacy. Ptolemy's model continued this tradition of growing complexity. To provide an accurate *predictive* model for projecting future motions of the known celestial bodies, an essential prerequisite for the practice of astrology, Ptolemy developed an elaborate system of multiple circular motions. The actual details of Ptolemy's system are of interest today mainly to historians. However, a few examples of the observational challenges, and the way they were answered, are instructive.

By observing the planets over the course of several months, it can easily be seen that they vary in brightness. This is difficult to accommodate within a philosophy which expects that the heavens are perfect and unchanging, unless the distance between the planet and the Earth changes

Retrograde planetary motion presents a difficulty for geocentric cosmological models

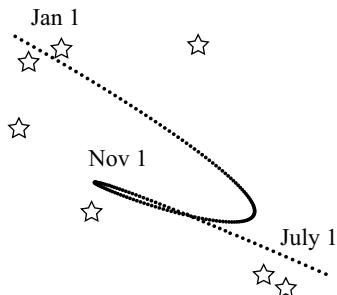


Fig. 2.3 The position of Mars relative to the background stars, plotted for the interval from July 1, 1988 to January 1, 1989. During that time Mars slows, stops, and reverses itself, travels backward, then reverses again and continues in the forward direction. This cosmic pirouette is known as retrograde motion.

Increasing complexity was needed to match increasingly detailed observations

with time. Another interesting planetary behavior is known as **retrograde motion**; this occurs when a planet reverses its usual direction with respect to the fixed stars and moves backwards for a while before resuming its forward motion. In addition to such directional changes, the speed with which a planet moves with respect to the background of fixed stars varies with time.

Like Eudoxus and Aristotle before him, Ptolemy was obliged to construct a hierarchical system of circles in order to account for the observations. The major circles, which carried the planets around the sky, were called the *deferents*. Superposed on each deferent was a smaller circle, the *epicycle*. With the addition of epicycles, the planets no longer executed strictly circular motion, although the net motion was still a sum of circular motions. Ptolemy shifted the center of the deferents away from the center of the Earth so as to account for the apparent changes in brightness and speeds of the planets. The net center of motion of each planet was also moved away from the center of the Earth, to a point called the *equant*. As viewed from the equant, the rate of rotation of the planet was constant. However, this new feature meant that the center of motion no longer corresponded with the supposed center of the universe.

The resulting model described planetary motions well, but in subsequent centuries it fell prey to the same failings as earlier cosmological systems: the accumulation of error over time, and improved observations. It became necessary to tinker further with the system, adding epicycles upon epicycles, the “wheels within wheels,” in an attempt to achieve the elusive perfection. Accuracy was obtained at the expense of simplicity, a fact that was not lost upon even adherents of the system. Alfonso, a 15th-century king of Castile and Leon, is said to have remarked upon learning the Ptolemaic system, “If the Lord Almighty had consulted me before embarking upon Creation, I should have recommended something simpler.” In retrospect, we can see how this increasingly elaborate and cumbersome construction continued to succeed. The true motions of the planets are not circular, but elliptical, and are centered upon the Sun, not the Earth. Nevertheless, any arbitrary closed curve can be approximated by a sequence of circles. But perfect accuracy requires an infinite number of circles, so ultimately the Ptolemaic system was bound to fail.

Ptolemy and his successors probably did not intend for their system to be taken literally, although ultimately its fate was to be taken all too literally. Their original purpose was a model that would serve as a mathematical tool to predict the positions of the planets. In that, the Ptolemaic system was quite successful for hundreds of years. It was eventually rejected not because it was inaccurate or incapable of correction, but because the heliocentric model proved to be much simpler. Moreover, the Ptolemaic model had no underlying, unifying predictive principle. If a new planet were discovered, the model could not describe its motion in advance, but only after many observations had been made to fit the required deferents and epicycles. The scientific method gives preference to the simpler theory with greater predictive power. The

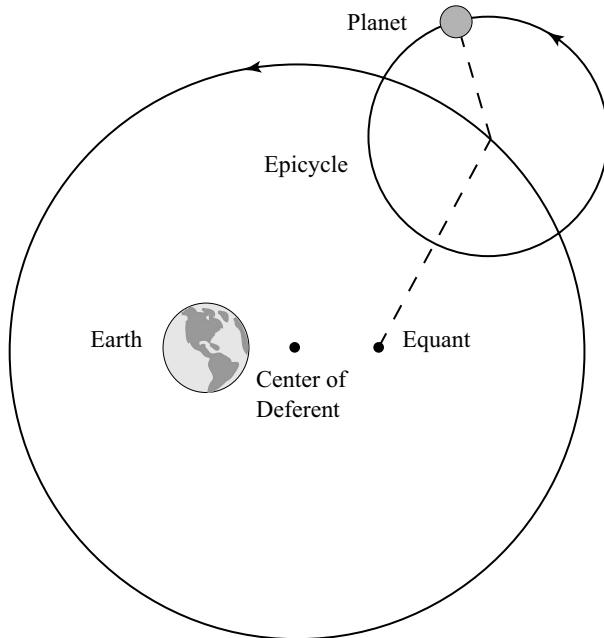


Fig. 2.4 Components of the Ptolemaic model for planetary motion. The planet moves on a small circle, the epicycle, which itself moves around the Earth on a larger circle called the deferent. The center of the motion is the equant.

heliocentric model that ultimately arose has taken us far, yet all models must be constantly tested by observations. Indeed, the Newtonian model of the solar system has a tiny but significant discrepancy in the orbit of Mercury, which eventually contributed to the acceptance of the general theory of relativity.

The Renaissance

With the decline of Greek culture, scientific cosmological modeling came to a halt. Greek learning was preserved by the Arabs, who added further observations to the growing volume of data and made additional refinements to the Ptolemaic system. Some Arab scholars were dissatisfied with Aristotelian physics and wrote detailed critiques of it, but no new theory arose in the Middle East. Aristotle's writings, along with further elaborations by his successors and by Ptolemy, were rediscovered in Europe at the beginning of the 13th century. The Greek/Ptolemaic cosmology eventually became incorporated into medieval European philosophy, with sufficient modifications to be compatible with Judaic and Christian theology. One important alteration was the change from a universe of infinite duration to one with a creation from nothing at a finite time in the past. The Earth remained at the center of the cosmos, although not because the Earth was considered to be an especially wonderful place. Indeed, in this cosmology the center of the Earth was the lowest, basest point of the cosmos, the location of Hell. The celestial realms were the domains of angels, with God beyond the outermost sphere. In this form, Thomas Aquinas and other medieval theologians

The Dark Ages

elevated the pagan Ptolemaic cosmology and Aristotelian physics into a cornerstone of Christian doctrine.

The supremacy of Aristotelian authority throughout the Middle Ages may well have occurred because, in essence, he told Europeans what they wanted to hear at the time. It was an authoritarian era, when the control of the Church in matters of belief was absolute, and dissent, whether in theology or science, was not tolerated. Aristotelian physics and especially his cosmology fitted the prevailing attitudes. It was believed that all that could be discovered had already been discovered. The search for new knowledge was regarded as a pointless enterprise, since Aristotle had anticipated and resolved all questions. Human curiosity could not be suppressed forever, however. The rediscovery of Greek scientific thought began a transformation in Europe that led eventually to the Renaissance. By the 1400s, every educated European was versed in Greek learning. Astronomy, which was more like what we would today call astrology, was one of the original liberal arts. For example, the English poet Geoffrey Chaucer wrote a treatise on the use of the astrolabe, an instrument for measuring the positions of stars. Educated Europeans also were well aware that the Earth was a sphere, and even knew its diameter to fairly good accuracy.

Given that most of the ancient Greek knowledge was well disseminated among the European elite of the 15th century, it is an interesting historical tidbit that, regardless of what some legends might claim, Christopher Columbus certainly was not waging a lonely battle against ignorance by contending that the Earth was spherical. On the contrary, Columbus had carried out his own erroneous calculation of the diameter of the Earth; he argued that it was a much *smaller* sphere than others believed, and maintained that the great Ocean was traversable by the small sailing ships of the era. In this case, conventional wisdom was correct and the supposed iconoclast was wrong. Others asserted, quite correctly, that a journey in a small sailboat across the distance proposed by Columbus was impossible. Columbus would have vanished, both from his countrymen and from history, had not an unknown (to Europeans of the time) continent intervened. The myth that Columbus was fighting the ignorant scholars of the time who insisted that the Earth was flat is pure fiction, apparently invented from whole cloth a few hundred years after his voyages and popularized by the writer Washington Irving. Columbus himself refused to accept that he had found a new land, believing to his dying day that he had discovered a route to Asia. Sometimes, it would seem, it is more important to be lucky than to be right.

The intellectual community of Europe in the 16th century was in a ferment. The increased level of literacy and education, the rediscovery of ancient scholarly works, and the development of printing raised the intellectual standards and dramatically altered the political climate. This new environment made possible such changes as the Reformation, which directly challenged the prevailing doctrinal authority of the time, the Roman Catholic Church. It is ironic, then, that the man who was

Columbus set out to prove that the Earth was small, not that it was spherical

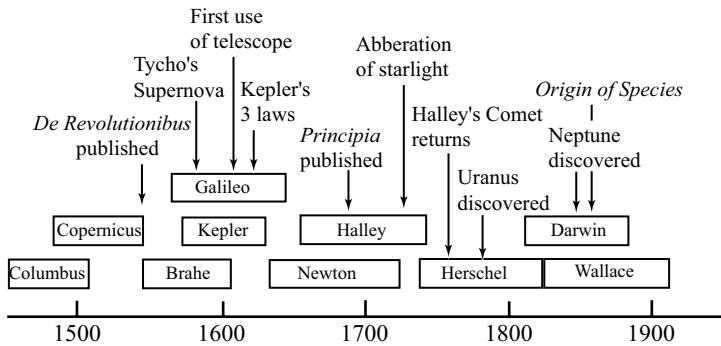


Fig. 2.5 A timeline for cosmological discovery prior to the 20th century.

to set into motion the coming cosmological revolution should have been a canon, a cathedral officer, in the Church. This man was Nicholas Copernicus.

Copernicus

Nicholas Copernicus is the Latinized name of the Polish scholar Mikolai Kopernik, who is credited with the introduction of the proposal that the Earth revolves around the Sun. This is called the **Copernican revolution**, and it was a revolution in more than one sense of the word: the revolution of the Earth, and a revolution in thought. Copernicus was not the first to propose such a *heliocentric*, or Sun-centered, system. Aristarchus of Samos had anticipated him by 1700 years; Copernicus apparently learned from one of his teachers about the work of Aristarchus. Copernicus, however, introduced his system into a world that was more receptive to new ideas, although it still was many years before heliocentrism was generally accepted. Indeed, Copernicus released his work *De revolutionibus orbium coelestium* (On the Revolution of Heavenly Spheres) for publication only near the end of his life. It appeared in 1543, and immediately created a sensation among the literate scholars of the day.

The Copernican revolution was the claim that the Earth is not the center of the universe

Why did Copernicus propose such a radical change? We can only speculate, as he left no explanation for his reasoning, but he apparently had several motivations. First, he was dissatisfied with the complexity of the Ptolemaic system. The continued addition of epicycles and eccentrics had made a mockery of the original goal of geometric purity in the celestial motions. Copernicus may well have hoped that by shifting the center of motion to the Sun he could restore the heavens to simple circular motion. He was also aware of the inaccuracies in the predictions of planetary positions, and must have expected that his model would make better forecasts. It also appears that he might have been attracted to the model by aesthetic considerations; where better to light the worlds than from the center of the universe?

The new theory had some immediate successes; for one, it explained the daily motions of the Sun and stars in terms of the simple rotation of

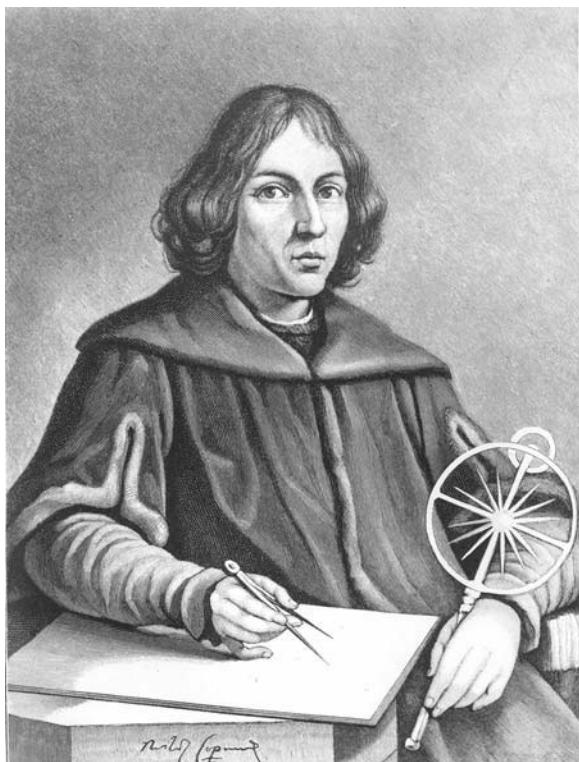


Fig. 2.6 Nicholas Copernicus (1473–1543), the Polish scholar whose Sun-centered model of the cosmos marks the beginning of modern astronomy. (Courtesy of Yerkes Observatory.)

A simple explanation for retrograde motion

the Earth. The seasonal changes in the patterns of the fixed stars was comprehensible as a consequence of the Earth's journey around the Sun, thus dispensing with the deferent that carried the fixed stars around the Earth in the Ptolemaic system. Even more prominently, it explained retrograde motion in a very natural way. The planets are like sprinters running around the lanes of a circular track. The innermost racers run faster, with outer racers lagging ever slower. We are the sprinter in the third lane out. As we overtake and pass the slower outer runners, they appear to move backwards with respect to the distant background, resuming their apparent forward motion after we are well past. Similarly, the inner runners are moving faster and pass us; as they turn the corner, we see them briefly move backwards. The Copernican system also made it possible to compute the relative spacing of the planets in their orbits. The two inner planets, Mercury and Venus, never travel far from the Sun in the sky. Simple geometry, combined with measurements of the angle of their maximum elongation away from the Sun, provides the size of their orbits, relative to that of the Earth. A similar, albeit slightly more complex, calculation gives the relative sizes of the orbits of Mars, Jupiter, and Saturn. In the Ptolemaic system the diameters of the various spheres were arbitrary, and were usually computed by assuming that they nested so as just to touch one another.

These advantages made the Copernican model the subject of much interest and discussion well before its formal release as a printed work. With regard to improved accuracy for observed planetary positions, however, the model failed. Copernicus placed the center of the cosmos at the Sun, but he still relied upon uniform circular motion. In the end, to fit the model to the known observations, Copernicus was forced to include many of the same complexities as the Ptolemaic system: equants, epicycles, and so forth. What Copernicus did not know was that circular motions would not suffice. Planets move on ellipses, not circles, but the true elliptical nature of planetary motions had yet to be discovered. Furthermore, parallax remained a sticking point for a heliocentric model, just as it had been for Aristarchus so many centuries before. Because stellar parallax had never been observed, Copernicus was obliged to expand greatly the size of the cosmos. He himself continued to regard it as finite, with a fixed sphere of stars removed to a great distance, but once the intellectual wall was breached and the heavens no longer hung close to the Earth, others grasped that the distances might be enormous, perhaps even infinite.

Although Copernicus did not produce a better cosmology, in the sense that Copernican predictions of celestial motions were not as accurate as those of the well-refined Ptolemaic model, his model did have an appealing simplicity. In one particular respect the Copernican model had a clear advantage over the Ptolemaic system: it made a *prediction*. In arranging the planets in their proper order from the Sun, Copernicus discovered that the inner planets moved faster than the outer ones. Thus, if a new planet were to be discovered farther from the Sun, it should be found to move more slowly than the known planets. However, he proposed no law to explain *why* the planets moved as they did; this explanation had to await the arrival of Newton.

For Copernicus, the inability of his model to make precise predictions of planetary positions meant failure, and may represent part of the reason that he did not publish his work until the end of his life. His book, which appeared well after his theories were already widely discussed, was a highly technical work, read by few. Why, then, was Copernicus so revolutionary? By abandoning the geocentric model, Copernicus struck at the philosophical underpinnings of the prevailing cosmology. In the Copernican system the Earth is not the center of the cosmos; it is just another planet. This development, with further elaboration, is now embodied in what is often called the **Copernican principle**, which, in its most elemental form, states that the Earth is not the center of the universe. This principle is the most valuable legacy of Copernicus.

The Copernican system was obviously a much more severe challenge to medieval theology than were any of the Greek models. Many passages in the Christian scriptures support the model of a stationary Earth, including the command by Joshua that the *Sun* should stand still. Belief in the Copernican system came to be regarded as heresy, and was suppressed by both the dominant Roman Catholic Church and the renegade Protestants. The Catholic Church still wielded formidable political power with

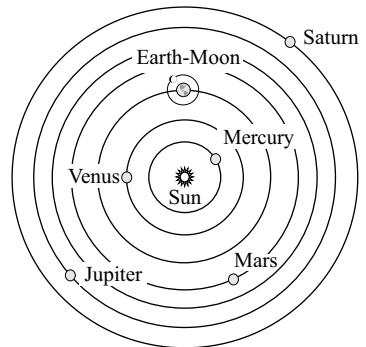


Fig. 2.7 The heliocentric model of Copernicus (not to scale). Copernicus arranged the planets in their correct order, and computed accurate relative spacings between them. The stars remained on a fixed sphere, now further removed so as to explain the lack of observable parallax.

The Copernican principle states that Earth is not located at a privileged position



Fig. 2.8 Tycho Brahe (1546–1601). Tycho's meticulous naked-eye observations of the heavens revealed the inadequacies of the Ptolemaic tables and provided the essential data that enabled Kepler to formulate the laws of planetary motion. (Courtesy of Yerkes Observatory.)

which to back its damnations, and at the time it was fighting the ultimate challenge to its authority, the Protestant Reformation. Dissension from accepted theology was thus especially dangerous. This alone was ample reason for the timid Copernicus to avoid publication as long as possible. At this he was quite successful; the page proofs for his book arrived as he lay dying. It may be that Copernicus developed an idea whose consequences ran away from him. He intended to save the phenomenon, to restore the Platonic purity of the circle, and to recreate the geometric beauty of the heavens as it was originally conceived. Instead he set in motion a revolution that would not be complete until both the cosmos, and the very foundations of physics, had been overturned.

Tycho Brahe

The intellectual climate of the Renaissance was receptive to the new Copernican ideas, but the most important driving force leading eventually to their adoption was increasing dissatisfaction with the Ptolemaic tables. With the development of the printing press, the tables were widely and accurately disseminated. Errors in the prediction of an important conjunction of planets by a few days could be blamed only upon the tables and not on the stars, nor on transcription errors. Although the telescope had not yet been invented, increasingly accurate observations made the faults of the Ptolemaic model all too apparent. This astro-

nomical trend reached its peak in the work of the Danish astronomer Tycho Brahe, the last of the great naked-eye observers.

Tycho is memorable both as a methodical scientific observer and as a remarkable personality. He was a member of the aristocracy, yet he devoted himself to the decidedly unaristocratic art of astronomy. In this pursuit, he benefited from his association with King Frederick II of Denmark, whose financial support enabled Tycho to build Uraniborg, a lavish observatory on an island just offshore from Copenhagen. Here Tycho lived the life of a self-indulgent nobleman while still devoting both his own efforts, and those of a considerable staff, to gathering his detailed observations of the heavens. Tycho's personality stood in marked contrast to his careful scientific work. He was a flamboyant and fiery man who sported a metal nose, the original having been cut off in a sword duel in his youth. He loved parties, which in his time were often lengthy binges involving much heavy drinking. He may have met his end as a result of such customs. Legend has it that he imbibed excessively at a royal banquet in Prague in 1601, but the protocol of the day prohibited guests from leaving the room when royalty was present. Tycho died shortly after this banquet, possibly as the result of a ruptured bladder.

It is easy to focus on such interesting details of Tycho's personal life, but he should be remembered instead for his exceptionally careful and systematic observations of celestial motions. Tycho repeated his measurements and used the additional information to estimate his errors, a revolutionary idea at the time. In this he was one of the first investigators who could be called a scientist, in the modern sense of the word. His amassed data provided a record of unprecedented accuracy and detail, and clearly showed the deficiencies in the Ptolemaic tables. Better observations do not simply destroy old theories; these observations were also accurate enough to allow Johannes Kepler finally to determine the correct planetary orbits, thus laying to rest forever the Ptolemaic system and establishing the basis for Newton's laws of motion.

In addition to his catalogue of accurate stellar and planetary positions, Tycho made several important discoveries. In 1572 he observed what was, at the time, an unbelievable sight, the sudden appearance of a new star in the constellation Cassiopeia. This was what we now call a supernova, a stellar explosion. When Tycho was unable to measure a parallax for this object, he realized that it could not be merely a brightening in the atmosphere of the Earth, but must belong to the realm of the fixed stars. This showed that the heavens were not immutable, a stunning revelation at the time. Tycho also demonstrated, again by means of parallax, that the orbit of a comet lay beyond that of the Moon. Until that time, comets had been believed to be vapors in the atmosphere of the Earth. Suddenly, the Aristotelian view of a perfect, changeless, unblemished heaven was untenable. New stars appeared and then disappeared. Unpredictable, rapidly moving comets belonged to the celestial realm. Indeed, the Aristotelian physics then accepted required that the crystalline spheres be real, physical entities; Aristotle

Tycho Brahe was the last great pretelescopic observer

Parallax is employed to judge celestial distances

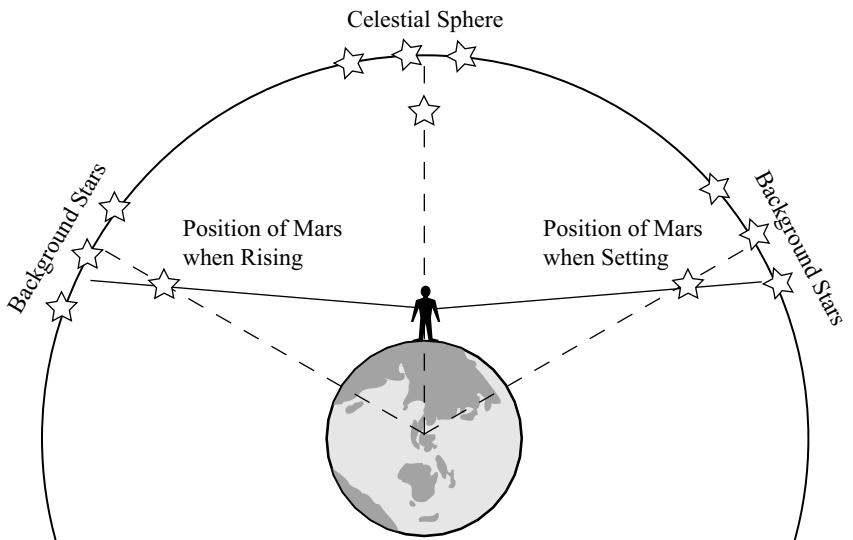


Fig. 2.9 Since Mars is closer than the celestial sphere, its position with respect to the background stars should shift over the course of a day. This is the diurnal parallax.

Tycho's attempt to perform a crucial experiment

believed that “nature abhors a vacuum” and thus he asserted that the spheres must fill all space. Yet now it seemed that the comets followed a path which must take them *through* the planetary spheres.

His suspicions about the Aristotelian model raised, Tycho began in 1582 an attempt to discriminate between the Ptolemaic and the Copernican models by measuring the distance to Mars when it was at its point of closest approach. He calculated that the diurnal parallax of Mars should be measurable if the solar system conformed to the Copernican model. The diurnal parallax is the change in apparent position produced by the change in the observer’s location due to the daily rotation of the Earth (Figure 2.9). At first Tycho was unable to measure any such parallax; in a later attempt with better instruments, he measured a negative angle, a nonsensical result. He decided that the culprit was refraction of the light by the atmosphere. He developed a table of corrections for this effect, attempted his measurement again, and obtained a credible result; at first he believed that he had succeeded at his goal. Tycho knew, however, the importance of checking and testing a result. He repeated his measurements with Jupiter and found, to his chagrin, the same parallax. Since Jupiter and Mars could not be at the same distance, he knew that it was his table of refractions that was incorrect, and consequently his results with Mars could not be accepted. We know now that Tycho’s program of planetary parallax measurements was doomed to failure from the beginning, because he was working from an inaccurate measurement of the size of the Earth’s orbit. From the time of Aristarchus until Kepler almost two millennia later, the distance from the Earth to the Sun had been underestimated by a factor of 20, rendering Tycho’s expected parallaxes far too large; he thought the result for Mars should be 5 minutes of arc, whereas the actual value is a minuscule 20 arcseconds, much

too small to be visible to the unaided eye.¹ The true scale of the solar system was much larger than anyone could fathom at the time. Tycho never realized why his project had failed, but he was honest enough to admit that his results were not valid.

Although primarily an observer, Tycho was not above trying his hand at cosmological modeling. Tycho was no Aristotelian; he knew particularly well the failings of the Ptolemaic system. Yet neither was he a Copernican. He ultimately rejected the heliocentric model because he was unable to detect stellar parallax. He knew that the lack of observable parallax could be explained by only two hypotheses: either the stars were so far away that their parallaxes were smaller than his measurement error, or else the Earth did not move. Tycho believed that the stars were near because he thought he was able to detect their apparent sizes. He did not realize that the finite disks of stars are an optical illusion, caused by the shifting of parcels of air in the Earth's atmosphere (stellar twinkling). If the stars had the sizes he measured, such great distances as were required by their lack of parallax implied them to be enormously large objects. Hence he concluded that the Earth could not be in motion. Tycho was a true scientist; he proposed a test of the heliocentric theory: the stellar parallax. The theory seemed to fail his test, so he rejected it. But even though he was not a Copernican, he did appreciate the simplicity of the heliocentric theory. Faced with conflicting observations and philosophical leanings, he proposed his own model in which the Sun and Moon revolved around the Earth, but everything else revolved around the Sun. In essence, he recreated the Copernican model, but shifted the center back to the Earth. Aside from differences in the frame of reference, the two systems were nearly equivalent. Like most compromises, however, Tycho's model pleased no one, except possibly himself.

Stellar parallax is an important prediction of the Copernican theory, and Tycho's objection was taken seriously. But the true distances to the stars are so great that Tycho could not possibly have detected any parallax without advanced telescope technology. If we were to shrink the radius of the Earth's orbit to a meter, the distance to the nearest star, Alpha Centauri, would be 274 kilometers (around 170 miles)! Measuring the parallax of this star amounts to determining the smallest angle of a triangle whose short side has a length of one meter and whose two long sides are 274,000 meters long. This angle works out to be less than an arcsecond, approximately one hundred times smaller than the unaided eye can resolve. It was not until 1838 that F. W. Bessel, F. G. W. Struve, and T. Henderson independently detected the parallaxes of the stars 61 Cygni and Vega, in the Northern Hemisphere, and Alpha Centauri, visible only from the Southern Hemisphere, thus proving once and for all the heliocentric model. Observations of stellar parallax retain their

Tycho's cosmological model

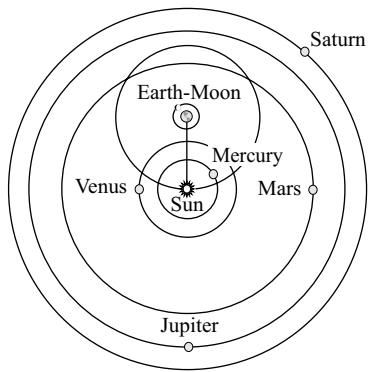
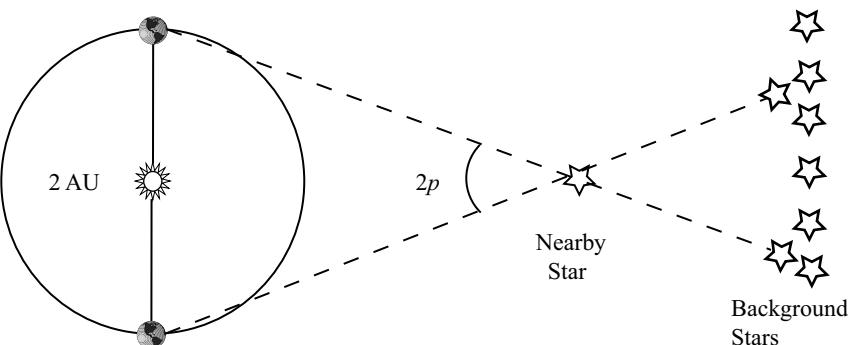


Fig. 2.10 Tycho Brahe's cosmological model. Earth remained the center of the cosmos and the Sun circled the Earth, but the other planets revolved around the Sun.

Stellar parallax finally detected

¹Distances on the sky are angles, which are measured in units of degrees, minutes, or seconds of arc. A minute is 1/60th of a degree, and a second is 1/60th of a minute. See Appendix B on units.

Fig. 2.11 Earth's annual orbital motion produces an apparent shift of a nearby star's position on the sky with respect to the background stars as our vantage point changes with the seasons. This shift is called parallax; observations of the parallax angle determine the distance to the star via triangulation. The base of the triangle is the diameter of the Earth's orbit; the mean distance between the Earth and the Sun is the Astronomical Unit (AU). The figure is not to scale; actual parallax angles are less than a second of arc.



cosmological importance even today, as they are still the fundamental basis for all stellar and galactic distance measurements.

Parallax is so difficult to observe for even the nearest stars that the first proof of the Earth's motion was indirect and came as late as 1728, more than a century after the deaths of Tycho and Kepler. The English astronomer James Bradley was attempting, unsuccessfully, to measure parallaxes when he noticed that *all* stars he observed showed a systematic shift with the seasons. At last the explanation came to him while he was boating; watching a vane turn with the winds, he realized that the Earth was traveling through a “wind” of starlight. An even better analogy is a sprint through the rain. If a pedestrian is caught outside without an umbrella in a sudden downpour, he must tilt his body forward in order that the newspaper he tries to hold over his head can be oriented perpendicular to the raindrops, even when the wind is perfectly still and the rain is falling straight down. The apparent direction of the source of the rain shifts because of the walker’s motion. The phenomenon discovered by Bradley is called the *aberration of starlight*.

The old model of the universe was disintegrating; yet there remained the task of building the new. Tycho’s observations did as much as anything to chip away the foundations of the prevailing cosmology, but his own attempt at a new cosmological model met with indifference. Clearly he was not the man who could create the new synthesis. It happened, however, that Tycho became embroiled in a dispute with King Christian, the sovereign of Denmark who ascended to the throne after the death of King Frederick, Tycho’s exceptionally generous benefactor. Tycho packed up his instruments and records in 1597 and moved from his private island off the coast of Denmark to central Europe. Tycho’s misfortune was the great fortune of science, for there he took a new assistant named Johannes Kepler.



Fig. 2.12 Johannes Kepler (1571–1630). Kepler's three laws of planetary orbits provided the first simple, predictive description of celestial motion. (Courtesy of Yerkes Observatory.)

Kepler

Tycho's schizophrenic cosmology was characteristic of a transitional era; the established model was rapidly failing, but its successor, the heliocentric model, had not yet been established. It fell to Johannes Kepler to develop the new paradigm. Kepler, a reserved Bohemian Protestant, came to Prague to work with the temperamental and outgoing Tycho in 1600, and set about interpreting his data. After Tycho's death in 1601, Kepler absconded with a vast collection of observational data. Its study occupied him for the rest of his life. Kepler first settled upon the objective of explaining the motion of Mars, a project suggested by Tycho, apparently because Mars shows the most irregularities in its motion. (We now know that this is because of its unusually eccentric orbit and its proximity to Earth.) He spent years considering all manner of epicycles. Nor did he find better luck from a different philosophical approach, in which he fitted the observations to his pet geometrical objects, a class of figures called Platonic solids. He felt that these objects had just as much right to be perfect as did a sphere, since they can be precisely surrounded by a sphere, but they yielded no improvement. Kepler did eventually hit upon a traditional Ptolemaic scheme that fit the observations better than any existing model of the time. He could

Kepler's search to understand the motions of the planets

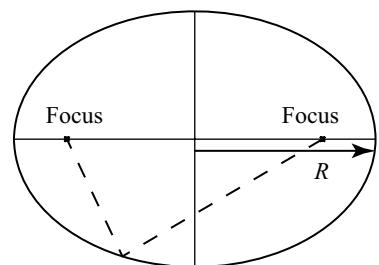


Fig. 2.13 An ellipse is the curve traced by a constant sum of distances (the dashed line) from two focus points. The semimajor axis, R , is indicated by the arrow.

have stopped at that point, and continued in his modest employment as a fairly successful man. But he was ruthless with himself and strictly intellectually honest. He was aware that none of his models, even his best, could describe the planetary motions to within Tycho's stated errors, and he was confident that Tycho had estimated his errors accurately. Therefore, Kepler struggled onward.

Finally, in 1604, he achieved success. Some inspiration caused him to abandon the ancient philosophical prejudices and to consider the motion of Mars as seen from the Sun. He found that he was able to fit the data to within the observational errors with an **ellipse** rather than a circle. The ellipse is the curve representing a constant sum of the distance from two fixed points, called *foci* (singular: focus). Because of its oval form, the ellipse has not a single diameter, but two perpendicular axes, the major (longer) axis and the minor (shorter) axis. The shape of an ellipse depends upon the separation of its foci. As the foci move further apart, the ellipse becomes increasingly elongated, or *eccentric*; conversely, a circle is a degenerate ellipse whose foci coincide. Thus an ellipse is really a generalization of the circle, so the ancients were not quite so far wrong after all. The eccentricity, or deviation from circularity, of the orbits of almost all the planets is very small; for the Earth's orbit, the major axis is a mere 0.014% longer than the minor axis. However, these relatively small differences from circular motion were more than sufficient to confuse astronomers for many centuries. Kepler discovered that the Sun was located at one focus of the ellipse. (The other focus is empty.) Each planet moved on its own elliptical path, with its own eccentricity. This insight was to unlock the secret of the heavens, although the work had only begun. It was not until 1621, after laborious calculations using the only mathematical tools available at the time, that Kepler finally arrived at his three laws of planetary orbits.

Kepler's three laws of planetary motion

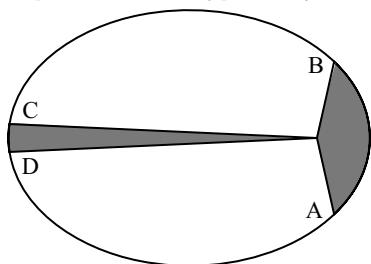


Fig. 2.14 The law of equal areas. A planet moves from *A* to *B* in the same time as it moves from *C* to *D*. The gray regions indicate the area swept out during the time to move from *A* to *B* or from *C* to *D*. These times are equal, as are the two indicated areas, according to Kepler's second law.

Kepler's first law: *Planets orbit the Sun in an ellipse, with the Sun at one focus.*

Kepler's second law: *The line from the Sun to the planet sweeps out an equal area in an equal time.* Thus planets move faster when they are nearer the Sun.

Kepler's third law: *The square of the period of the orbit is equal to the cube of the semimajor axis (half the long axis) of the ellipse.*

If the period, symbolized by P , is measured in years, and the size of the semimajor axis of the ellipse, R , is measured in terms of the *astronomical unit*, where the AU is defined as the mean distance of the Earth from the Sun, then this law can be expressed mathematically as

$$P^2 = R^3. \quad (2.1)$$

Kepler had strong mystical leanings and always hoped to find deep meaning in the cosmos. His third law, often called the *harmonic law*, was probably the most personally satisfying discovery of his life. Kepler went so far as to assign musical notes to the planets, based upon his third

law. Today the mathematical beauty of the harmonic law is understood to be a direct consequence of more fundamental, and perhaps even more beautiful, laws of physics. In Kepler's time, however, this achievement was a great triumph. It is fair to say that Kepler was the first to hear the true music of the spheres.

With Kepler's laws in place, simplicity swept away complexity. There was no need for circular motion; the Copernican system, freed of its epicycles, finally revealed the elegant simplicity of the travels of the planets around the Sun. Now it could be shown that the new model agreed with observations to a far better precision than even the carefully elaborated Ptolemaic system. The complexities of the geocentric systems were due not only to their inappropriate frame of reference but, in retrospect, to the impossibility of fitting an ellipse with any finite sequence of circles. The data were forcing a change, but the idea of the primacy of circular motion was so strong in European thought that the correct solution could not have been seen. It was Kepler's great achievement that he was able to break through this mindset. And it was not so much that the old theory was demolished as it was a crystallization of what was already known, now seen in a new light. The old theory had reached the end of its possibilities.

Kepler's laws provide a correct mathematical description of planetary motion. Unlike the Ptolemaic model, the Keplerian model has considerable predictive power. If a new planet were discovered, not only could we predict whether it would orbit faster or slower around the Sun but, from only a few observations to determine the length of the semimajor axis of the orbit, we could predict the period of that orbit. However, Kepler's laws alone do not provide much insight into why the motion should occur as it does. Kepler recognized that the third law provides a clue. If planets orbit more slowly the greater their distance from the Sun, then their motion must be related to some influence from the Sun. Sunlight also diminishes with distance from the Sun, so perhaps there is some force emanating from the Sun that sweeps the planets along in their orbit; this force must decrease with distance, just as does the intensity of sunlight. Unfortunately, Kepler still labored in the shadow of Aristotelian mechanics. Kepler lacked the proper definition of inertial (natural) motion, so he was not quite able to grasp the law of gravitation; the correct formulation had to await the arrival of Newton. Perhaps it is too much to expect a single individual to do more than to overthrow the cosmology accepted for two thousand years.

Kepler was a quiet and unassuming man who might not have seemed destined for the greatness he achieved. He was not highly regarded in his day, yet he was persistent, mathematically gifted, and intellectually honest. While he never completely abandoned his philosophical prejudices, continuing to think about his Platonic solids even after his success with ellipses, he was able to put them aside rather than allow them to twist his theories away from their observational roots. His achievements are eloquently summarized by Kepler himself, in his own epitaph. The orig-

The concept of the primacy of circular motion must be discarded

Kepler's Laws have predictive power

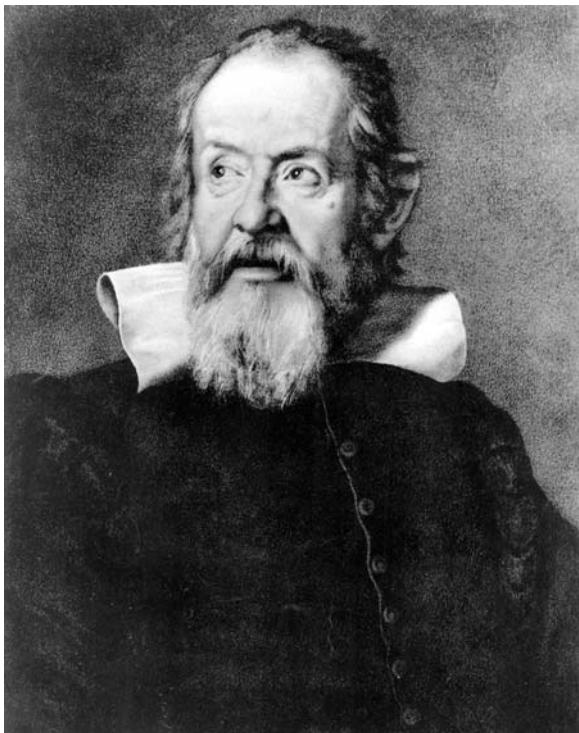


Fig. 2.15 Galileo Galilei (1564–1642), ardent champion of the heliocentric model. (Courtesy of Yerkes Observatory.)

inal was written in Latin, the scholarly language of the day; an English translation is:

*I measured the heavens, now I measure the shadows,
Skyward was the mind, the body rests in the earth.*

Galileo

Kepler was the scientist who discovered the mathematical laws of the celestial motions, and it was he who made the bold leap from circles to ellipses that finally vindicated the Copernican heliocentric system. Yet the name most popularly associated with the championing of this new worldview is that of the Italian astronomer Galileo Galilei. Galileo was one of the great Renaissance scientists. He made significant contributions in many areas of research, although he is most remembered for his astronomical discoveries, which he made by putting the newly invented telescope to its first celestial use. It is often believed that Galileo invented the telescope, a misconception common even during his lifetime and one that Galileo himself made no attempt to dispel. However, credit for the invention of the telescope is usually assigned to Hans Lippershey, a Dutch lens grinder, although earlier lens makers may have discovered the basic principles. In any case, as soon as Galileo heard of this new

instrument in 1609, he immediately built one and turned it toward the sky.²

One of Galileo's first observations was of craters and mountains on the Moon. This showed that the Moon was not a smooth sphere, but was a world with its own detail, much like the Earth. He also turned his telescope to the Sun. He did not discover sunspots (they had been, and still can be, observed by the unaided eye at sunrise or sunset), but he was the first to conclude correctly that the spots were associated with the Sun itself and were not foreground objects. Galileo also recognized that the Sun carried the spots around as it rotated on its own axis; this enabled him to estimate the rotation rate of the Sun. Observations such as these pounded away at the Aristotelian concept of the perfection of celestial bodies. As Tycho had discovered around the same era, the skies were not the abode of perfect, immutable objects. The Earthly and celestial realms were not distinct, but might obey the same laws and be made of the same substances.

Galileo made another surprising discovery when he turned his telescope toward the Milky Way, which to the unaided eye appears only as a diffuse glow spanning the sky. He resolved the glow into a myriad of stars too faint to see without the new device. But if these stars were too dim to see, while others were visible without the aid of the telescope, how could they reside upon the same crystalline sphere, as required by the ancient cosmology? Under magnification, the new, faint stars had the same apparent size as all the others. This suggested that the apparent disks seen by earlier observers, including Tycho, were an illusion. Even today, a sweep through the Milky Way with a simple pair of binoculars gives a distinct sensation of vast depth to the skies. The Copernican model and the lack of observable parallax required the stars to be at a great distance; the telescope made such a heresy believable.

Although the stars remain unresolved points even to modern instruments, Galileo found that the planets *did* present disks to his telescope; in fact, Venus went through phases, and its phases accounted for some of its dramatic changes in brightness. The gibbous and full phases of Venus observed by Galileo could not be explained by the Ptolemaic model, which could produce only crescent and new phases. The Ptolemaic model made a testable prediction about the phases of Venus, which it failed when the observation was made. The Copernican system, on the other hand, predicted a full range of phases; hence Galileo's observations are an example of a crucial experiment, providing strong evidence in favor of the heliocentric model.³

Galileo's telescopic observations

The phases of Venus constitute a crucial experiment

²Like many new technologies before and since, the initial applications of the telescope were for military purposes. Galileo demonstrated the military possibilities to the local authorities in Padua, impressing them sufficiently that they provided him with funding and status. After this, Galileo became the first to apply the telescope to scientific inquiry.

³Tycho's cosmological model would also produce phases in Venus as observed by Galileo. A crucial experiment to distinguish Tycho's model from the Copernican would be to determine whether or not the Earth moves.

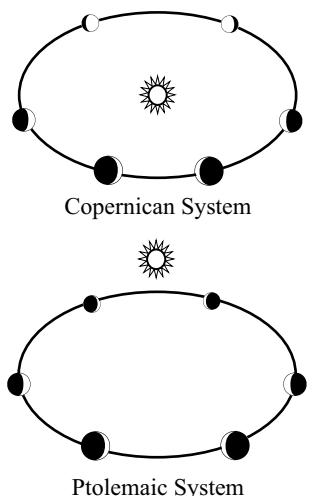


Fig. 2.16 Galileo's observations of Venus revealed a full ensemble of phases, from crescent to full. This is consistent with the Copernican model in which Venus circles the Sun, but not with the Ptolemaic model in which Venus always lies between the Earth and the Sun.

Galileo's studies of motion

Perhaps Galileo's most dramatic observation was that Jupiter commands its own miniature system. Galileo discovered the four largest moons of Jupiter, still known as the *Galilean moons*. It is one thing to observe new details of known objects; far more sensational is to discover completely new objects. Galileo's careful charting of the motions of these objects demonstrated unequivocally that the moons orbited Jupiter. The Earth was not the only center of motion, refuting one of the basic tenets of Aristotelian cosmology.

The impact of Galileo's findings was widespread. When he wrote of his observations in his book *The Starry Messenger*, he wrote in Italian rather than in the Latin of scholars, so that everyone could read about his discoveries. Soon many people were turning telescopes skyward to share in these new wonders.

Although Galileo began his career teaching the standard Ptolemaic model, he apparently was never satisfied with Aristotelian cosmology. He had little patience with his fellow scholastics, who unquestioningly repeated Aristotle's laws of physics. Galileo was not content to accept the word of even so venerated an authority as Aristotle, and often put the Aristotelian precepts to the test. When his astronomical observations converted him completely to the Copernican model, he was faced with the problem of reconciling his findings with physics. Aristotle's physics explicitly denied the motion of the Earth, which seemed to be perfectly consistent with the observations of our senses. Yet the skies supported the Copernican model. How was physics to be modified to explain this apparent contradiction?

Fortunately, Galileo had devoted much of his career to the physics of mechanics. In particular, he was intrigued by the motion of falling bodies. Aristotle held that the rate of fall depends upon the composition of the falling body, and of the medium through which the body fell. Galileo recognized that this idea could be tested, as indeed several other scholars of the time had done. He carried out his own experiments (none of which, apparently, involved dropping any objects from the Leaning Tower of Pisa), and made measurements in support of his conclusion that all objects fall at the same rate, contrary to the Aristotelian claim. But the limitations of the technology of his time forced him to appeal for many of his arguments to **thought experiments**, that is, mental experiments that could, in principle, be performed if the technology were available. As an example, consider a stone falling from a height. Now imaging cutting the same stone into two equal pieces, then dropping them together. Would the severed halves fall at different rates from the whole? What if the two pieces were connected by a short string? It should be clear that a boulder will not suddenly fall at a different rate if a crack appears in it. From such reasoning, Galileo concluded that all objects must fall at the same rate in a vacuum. This important observation, that in the absence of air resistance or other complicating factors, all objects fall at the same rate in a gravitational field, is now called the *equivalence principle*; Galileo was one of the first to articulate it clearly. Yet even Galileo could not have realized how profound was

this observation, as much later it became the basis of general relativity; more immediately, it formed a foundation of Newton's theory of gravity.

A key rule of mechanics, with which Galileo struggled, is the law of inertia. Galileo's knowledge of contemporary experimental results, plus his own experiments with pendula and with balls rolling on an inclined plane, convinced him that impetus was not lost, but was *conserved* in freely moving bodies. Hence not only does an object at rest remain so unless a force acts upon it, but a body in motion in a straight line remains in that motion unless a force acts. The essential break from Aristotelian mechanics to modern mechanics is to recognize that force is responsible not for motion, but for *changes* in motion. From this realization, the relativity of uniform motion follows. Galileo understood the experimental fact that if everything is moving together uniformly, such as the furniture and lamps in the interior of a moving ship, then it will seem no different from when the ship was at dock. To take a more modern example, imagine a trip on a supersonic passenger aircraft, such as the Concorde that made transatlantic flights for several decades. At dinner the flight attendant pours coffee normally. Flying faster than the speed of sound, a passenger feels no more sensation of speed than is felt while sitting in his living room. This leads to the conclusion that constant-velocity motion is not necessarily perceptible if the observer and his surroundings are moving together; hence the Earth could be moving through space, yet this may not be directly noticeable by the humans moving along with it. This was the critical conceptual breakthrough that made the heliocentric model plausible. However, Galileo never completely worked out the laws of motion that would replace those of Aristotle. That task fell to Isaac Newton.

Galileo summarized his cosmological conclusions in 1632 in a new book *Dialogues Concerning Two Chief World Systems*, in which he showed how his discoveries supported the Copernican system. The book caused a sensation throughout educated Europe and paved the way for the new paradigm of the universe. It also set the stage for Galileo's later troubles with the Church. His outspoken advocacy of the Copernican model had earlier discomfited Church authorities, and this new book provided further provocation. One of his political missteps was to place the defense of the Aristotelian cosmology into the mouth of Simplicio, an obvious fool. Galileo was brought to trial for heresy in 1633, was forced to recant his scientific beliefs, and was confined to his home for the rest of his life. Only in 1980 did the ecclesiastical authorities finally exonerate him.

Galileo was a vain, arrogant man; in the end, he came to regard himself as much of an authority as Aristotle had considered himself. He deliberately provoked the Church and was actually given an unusually light penalty at his celebrated trial, partly due to his fame and partly because of his advanced age and infirmity at the time he was brought before the Inquisition. Galileo certainly promoted himself and was not above claiming credit, or allowing credit to be assigned to him, for nearly every discovery in astronomy during his lifetime. Despite such character

The relativity of uniform motion

failings, however, he was an important figure in the history of science. He was one of the first to understand fully how critical is the role of experiment. Both he and Kepler realized that *data*, not our philosophical wishes, must be the final arbiter of science. One modern school of thought in the philosophy of science holds that great discoveries are more the products of an era than of individual genius. If Galileo had not made his discoveries, someone else would have done so. There is probably much truth to this idea, as it is clear from history that important discoveries are often made simultaneously and independently by more than one researcher. Yet there must be some due given to individuals. Perhaps it is the combination of the right person at the right time. Kepler and Galileo were the right people at the right time; between them they irrevocably changed our view of the world.

Chapter Summary

The first attempt to construct a systematic cosmology that was grounded in physical theory was the model of Aristotle. Aristotle developed a theory of motion and defined the concepts of natural motion and force. In Aristotle's view, the Earth was the center of the universe and the center of all natural motions. Motions on the Earth were linear and finite, while the heavenly bodies executed perfect circles eternally. The stars and planets were composed of a perfect element called ether, whereas Earthly objects were made up of varying combinations of the four ancient elements of earth, air, fire, and water; a body's motion was a consequence of its composition. Although our modern definitions of these concepts are quite different from Aristotle's, natural motion and force remain fundamental to our understanding of the structure and evolution of the universe. Aristotle's Earth-centered worldview was later embodied in the detailed model of Ptolemy, with its deferents, epicycles, and eccentrics designed to predict the complicated celestial motions of the planets while still requiring motion in the heavens to be built upon circles.

During the Renaissance, humanity's cosmological model changed dramatically. Copernicus developed a Sun-centered model of the heavens that gained rapid ascendancy in Renaissance Europe. Tycho Brahe's detailed naked-eye observations of the heavens provided the data that Kepler used to derive his laws of planetary motion. Kepler's laws of planetary motion made it possible for the first time for humans to understand the paths of the

wanderers across the sky. These laws were among the greatest quantitative achievements of the Renaissance.

Galileo, a contemporary of Kepler, was the first to make serious scientific use of the telescope, an instrument which provided observations that challenged the Ptolemaic model of the heavens. Galileo observed craters on the Moon, demonstrating that it was not a perfect, smooth sphere. He found that the Milky Way was not a solid band of light but was filled with myriad stars, too small to be resolved by the unaided eye. A key observation was that Venus went through a full cycle of phases, just like the Moon; this was impossible in the Ptolemaic model but was required by the Copernican model. One of Galileo's most important discoveries was of the four largest satellites of Jupiter. These bodies demonstrated that the Earth was not the only center of motion in the universe, thus refuting one of the important tenets of Ptolemaic–Aristotelian cosmology and physics.

Galileo also studied mechanics. From direct observation and careful reasoning, he was able to arrive at the conclusion that all bodies fall at the same rate, if air resistance is negligible. This principle, now called the *equivalence principle*, is one of the foundations of the general theory of relativity. Galileo also realized that motion might not be easily detectable by observers partaking of that motion. This was an important prerequisite to the work of Isaac Newton, who would later develop the fundamental laws of physics and gravitation that govern the universe under most conditions.

Key Term Definitions

geocentric Taking the Earth to be the center, for example of the solar system.

mechanics The science of motion.

force That which produces an acceleration.

inertia That property of an object which resists changes in its state of motion.

heliocentric Taking the Sun to be the center, for example of the solar system.

parallax The apparent shift in the position of a celestial object, such as a star, due to the changing vantage point of the observer. Astronomical parallax can be caused by phenomena such as the orbital motion of the Earth, or its daily rotation (*diurnal parallax*).

retrograde motion The apparent reversal in the motion of a planet across the sky relative to the background stars, caused by the Earth passing the planet or being passed by it.

Copernican revolution The revolution in thought resulting from the acceptance of the heliocentric model of the solar system.

Copernican principle The principle that the Earth is not the center of the universe.

ellipse A geometric figure generated by keeping the sum of the distance from two fixed points (the foci) constant.

Kepler's laws The three laws of planetary motion discovered by Johannes Kepler.

thought experiment An experiment that could be performed in principle but might be very difficult in practice, and whose outcome can be predicted by pure logic. Often used to develop the consequences of a theory, so that more practical phenomena can be predicted and put to actual experimental tests.

Review Questions

- (2.1) From what evidence did the ancient Greeks (and others) conclude that the Earth was immobile?
- (2.2) From what evidence did the ancient Greeks deduce that the Earth was a sphere?
- (2.3) Why did Eudoxus demand spherical motions for the planets? What were the consequences for his model of this assumption?
- (2.4) According to Aristotle, what caused motion on the Earth? In the heavens? What type of motion was appropriate to each realm?
- (2.5) What was the impetus theory of motion?
- (2.6) [More challenging.] While stationed on the planet Zorlo, you decide to replicate the calculation of Aristarchus for the Earth and the Sun. Zorlo's moon, Crastig, completes one revolution (360 degrees) in 42 Zorlo days. You observe that Crastig requires 20.985 days from third to first quarter. What is the ratio of the distance from Zorlo to its moon Crastig, to the distance from Zorlo to its sun? (Hints: first compute the number of degrees traveled by Crastig in one day. From Figure 2.2, note that the desired ratio of distances is given by the cosine of the angle α .)
- (2.7) How did Ptolemy account for the retrograde motion of the planets?
- (2.8) Describe two major weaknesses of the Ptolemaic model of planetary motions.
- (2.9) The imagery of Hell existing down below and Heaven having a location above the clouds is still common, at least metaphorically. How is this connected to medieval European cosmology?
- (2.10) Was the original Copernican model simpler than the Ptolemaic? What phenomena were more easily explained by the Copernican theory than the Ptolemaic? What is the most valuable legacy of Copernicus?
- (2.11) What Aristotelian belief did the observations of Tycho Brahe most seriously challenge? Why did Tycho reject the Copernican model?
- (2.12) What experiments did Tycho Brahe perform to test the Ptolemaic and Copernican cosmological models?

- (2.13) A new asteroid orbits the sun at a mean distance of 40 AU. What is the period of its orbit in Earth years? Does the answer depend on how elliptical the orbit is?
- (2.14) Describe three observations of Galileo which supported the Copernican model. State also why they falsified the Ptolemaic/Aristotelian system.
- (2.15) In what way did Galileo's observations on the properties of motion disagree with Aristotelian mechanics?

Newton's Machine

3

Key Terms:

I think Isaac Newton is doing most of the driving right now.

Astronaut Bill Anders, aboard *Apollo 8* during its return from lunar orbit.

- **Newton's first law**
- **uniform motion**
- **force**
- **acceleration**
- **Newton's second law**
- **mass**
- **conservation of momentum**
- **vector**
- **Newton's third law**
- **universal gravitation**
- **gravitational constant**
- **weight**
- **radioactive dating**

Isaac Newton

If modern physics and cosmology can be assigned a birthday, it would be that of Isaac Newton. Born prematurely in Lincolnshire, England, on Christmas Day of 1642,¹ according to the calendar then in use in England, the infant Newton barely survived. His father had died before his birth; when his mother remarried a few years afterward, he was given over primarily to the care of his maternal grandmother. He distinguished himself scholastically even as a child, and his family decided that he should enter a university; he began his undergraduate study at Cambridge University in 1661. Newton set about the study of mechanics, a science which at that time was still dominated by the theories of Aristotle. The University was not immune to new ideas, however, and Newton also acquainted himself with the more recent work of Kepler, Galileo, and the French philosopher René Descartes.

Newton left Cambridge in 1665 when the university closed during an epidemic of the plague. Returning home, he began the independent research that was to revolutionize science. During his 18 months in Lincolnshire, Newton developed the mathematical science of calculus, performed experiments in optics, the science of light, and carried out his initial derivations of the laws of mechanics and gravity. In 1669 Newton was named Lucasian Professor of Mathematics at Trinity College, Cambridge, in recognition of his accomplishments with calculus; he held this position for the rest of his scientific career. In 1672 Newton was elected to the Royal Society of London on the basis of his work in optics, particularly his invention of the reflecting telescope, a device which uses a mirror rather than a lens to focus light; this is still the basic design of all large astronomical telescopes.

Newton was an embodiment of the eccentric genius. In addition to his work in physics, he dabbled in alchemy and theology; he considered

The life of Isaac Newton

¹This corresponds to January 4, 1643 on our current Gregorian calendar.

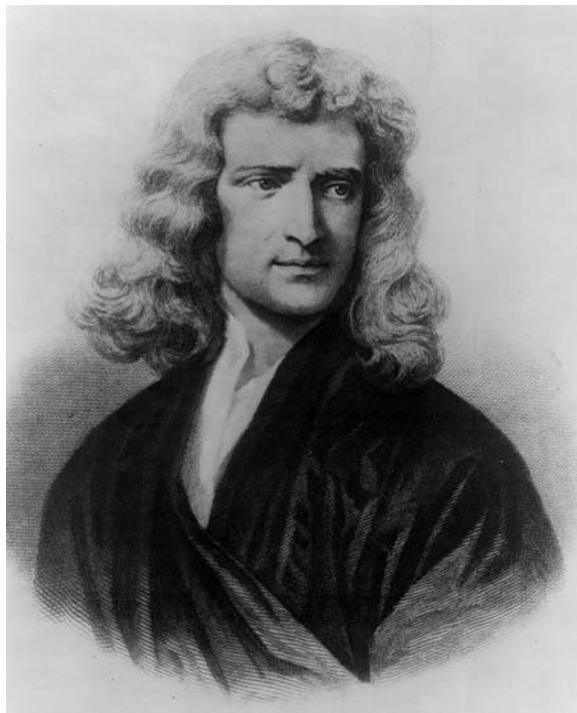


Fig. 3.1 Isaac Newton (1643–1727). His master work, the *Principia*, established the science of mechanics and provided the law of gravitation that explained Kepler's laws. (Courtesy of Yerkes Observatory.)

his efforts in those fields to be every bit as important as his physics, although today they are regarded to be at best of no consequence, and at worst completely wrong. Newton shared with Aristotle and Galileo the conviction that he was always right. He once ended a friendship because the friend dared to disagree with Newton's interpretation of the Old Testament's Book of Daniel. His pathological personality almost denied the world the benefit of his insights. Newton was neurotically protective of his privacy, becoming greatly disturbed when his first published work, a report of his discoveries in optics, drew international attention to him. The paper was important; Newton demonstrated that a prism separated light into colors because the index of refraction, a measure of the bending of light in a medium, differed for each color. He proved that white light was a mixture of these colors by showing that a second prism could recombine the spectrum into white light. It had previously been believed that a prism somehow manufactured the colored light internally. Although the paper was generally well received, this early optical work dragged Newton into a dispute with his chief rival, Robert Hooke. Hooke attacked the paper viciously because it made some sweeping statements about Newton's corpuscular theory of light, on the shaky basis of some rather crude experiments. Stung, Newton withdrew for a while even further into his shell. His full optical researches were not published until after Hooke's death.

Robert Hooke made important scientific contributions of his own; among other things, he discovered the rotation of Jupiter. Yet he is

also remembered as a thorn in Newton's side, an egotist who claimed for years that Newton had stolen "his" theory of gravitation. In fact, Hooke and some other intellectuals of his day had independently arrived at the hypothesis that gravity obeyed an inverse square law. (Kepler himself had suspected that the Sun's influence over the planets obeyed such a relation, so the idea was hardly new.) In January of 1684, Hooke boasted to Edmund Halley and Christopher Wren that he could prove this assertion easily, but he failed to produce a demonstration after several months. Later that year Halley, in Cambridge for other reasons, stopped to see Newton. Halley asked Newton what would be the orbit of a planet obeying an inverse square law of gravity, and Newton replied immediately that it would follow an ellipse. When the astonished Halley asked Newton how he knew this, Newton replied, "I have calculated it." Halley requested to see the work, but Newton, rummaging through his stacks of papers, claimed he could not find it. (Most likely, the proof was incomplete and contained an error, and Newton did not wish to expose himself to criticism.) Although Newton had carried out the basic calculations years earlier, he had delayed publishing his findings because he had difficulty in proving an important result in his theory of gravity. Now, goaded by Halley's request, Newton turned again to the problem. Over the following three months, Newton worked out the proof in detail and sent a copy to Halley, who immediately urged Newton to publish a full description of his work. Two years later, Newton delivered a manuscript that laid down the fundamental laws of mechanics and gravitation, laws that are still today the basis of mechanics in the usual limiting case of speeds that are not too extreme and gravitational fields that are not too strong.

Newton published this great work, the *Philosophiae Naturalis Principia Mathematica* (The Mathematical Principles of Natural Philosophy) in 1687. This book, issued with the imprimatur of the Royal Society, is usually known simply as the *Principia*. Halley paid for the publication from his own pocket; without the intervention and support of Halley, Newton's discoveries might never have reached the world. Had such a calamity occurred, at best the progress of science would likely have been much delayed. The laws of mechanics and gravity might well have trickled out slowly, attributed piecemeal to the work of others, rather than emerging, as they did, as a unifying whole. Edmund Halley, who had the patience to remain Newton's friend for years despite Newton's tantrums and quirks, might have been one of the few who could have persuaded him to publish.

Newton's scientific career ended only a few years after the *Principia* appeared. In 1693 he suffered an unmistakable mental breakdown, possibly at least partially due to years of exposure to mercury, a very toxic heavy metal, during his alchemical studies. He recovered, but never made any further contributions to science. By then, however, Newton's renown was so great that the British government arranged for him to receive a comfortable position; he spent the last thirty years of his life managing the British Mint, an office which brought his considerable ec-

The Principia was Newton's masterpiece

centricities into the public eye. He became a common object of ridicule, lampooned in popular plays and pamphlets as a pompous, overblown martinet. History has been more generous. Newton is now recognized as one of the greatest scientists in history, and the *Principia* as possibly the greatest scientific work ever published.

Newton's laws

The science of Newtonian mechanics, as elaborated in the *Principia*, is summarized in Newton's three laws of motion. These three laws can be stated quite briefly and in simple language, yet they are of overarching importance, transcending this apparent simplicity.

Newton's first law

The first law of motion returns to the question of natural motion, or how objects move if left on their own. Aristotle believed that things moved only if acted upon by a force, that is, a push or a pull exerted by one object upon another. An arrow flies only because it is pushed along by the air through which it moves. Otherwise, all things move to their proper location within the cosmos. This implicitly assumes that there is a universal standard of rest, relative to which everything moves. For Aristotle, this standard of rest was the center of the immobile Earth, the center of the universe. By Newton's time, however, scientists and philosophers were well aware that Aristotelian mechanics was incompatible with the developing cosmological model. According to the Copernican system, the Earth orbits the Sun and rotates on its axis; this implies that the Earth is executing stupendous motions, yet there is no obvious application of a force causing it to move, nor any sensation of motion for its inhabitants. To explain these facts, Galileo developed the idea that motion is *relative*: if all things move uniformly together, sharing in a common motion, there is no discernible effect. The Earth's motion through space is imperceptible because we take part in that motion. This suggests that there is no absolute standard of rest; the state of rest is relative. We can retain the aspect of Aristotelian physics which asserts that an object at rest will remain at rest unless acted upon by some force. But if the meaning of rest is relative, then motion cannot by itself require the continual application of a force. Consider two persons, each moving uniformly with respect to the other. Each feels himself to be at rest, and no force is required to remain so. Each feels no force, even though the two are moving with respect to one other. To initiate some other motion does require a force. It is not motion *per se* that requires a force, but a *change* in the state of motion. Hence we arrive at:

The relativity of uniform motion means there is no state of absolute rest

Newton's first law of motion: *A body at rest or in a state of uniform motion will remain at rest or in uniform motion, unless acted upon by a net external force.*

In this law, also called the *law of inertia*, Newton grasped what others had failed to see, that not only would a body at rest remain in that

state, but a body in **uniform motion**, that is, traveling in a straight line with constant speed, would also persist in that state unless a force acts. This law also clarifies and defines what is meant by a **force**: a force is that which causes a body at rest or in uniform motion to change its state. Note that the first part of the statement is really just a special case of the second part, since a body at rest has a velocity of exactly zero, which surely is uniform. In the absence of force, a body in uniform motion will remain in its uniform motion forever.

We often have difficulty in grasping intuitively Newton's first law because the motions we commonly experience are always affected by forces. The arrow, flying through the sky, is slowed down by the act of *pushing* the air out of its path. The force of air does not keep the arrow flying; to the contrary, it is the force due to air resistance that eventually brings it to a halt. Similarly, an automobile will come to a stop if the engine shuts off. This is a result of the resistive force exerted on the car's tires by the ground, a force we call *friction*. In the absence of friction an automobile, once started, would travel down a straight road without the need for any motive power whatsoever. More realistically, if the friction between the tires and the road is reduced, for example by driving on glare ice, the driver will quickly discover the difference between Aristotelian and Newtonian physics. Regrettably, uniform straight-line motion will continue until acted upon by the force of the collision with the tree.

Thus we conclude that a force is required to produce a change in velocity, where in physics velocity is defined as speed *and* direction. A change in velocity means a change with respect to time of speed or direction or both. This is an **acceleration**, and mathematically it is expressed as the change in velocity per unit time. But what is the relationship between force and acceleration? Is force simply equal to acceleration, or is it more complicated? We know that our arms can exert a force, for instance when throwing a ball. Presumably there is only so much push we can exert upon the ball with our arms, so the force we can produce is limited. And we know that the same force, when exerted upon a hollow rubber ball, produces a much greater velocity than when exerted upon a bowling ball. Therefore, the amount of acceleration generated by a force is linked with how massive something is. The exact statement of this idea is:

Newton's second law of motion: *The acceleration of an object is equal to the net force applied to it, divided by its mass.*

Acceleration defined

Newton's second law

Mathematically, this law can be expressed in the form

$$\mathbf{F} = m\mathbf{a}, \quad (3.1)$$

where **F** is the symbol for the force, including both its magnitude and its direction, **m** is the mass, and **a** is the acceleration, also with magnitude and direction. This simple law contains most of the science of mechanics. The force that appears in equation (3.1) is the *net* force, the sum of all forces acting upon the body. If you pull a wagon over a rough surface,

Newton's first law is not obvious in everyday motion

the horizontal forces on the wagon are your pull in one direction, and friction, which occurs whenever one object moves over another, in the other direction. The net horizontal force is the sum of these two forces, and this net force determines how successfully you accelerate the wagon. When you are pulling the wagon at a constant speed in a fixed direction, the force you exert is exactly the same in magnitude as, and opposite to the direction of, the frictional force, so the net force is zero, as Newton's first law requires.

A definition for mass as inertia

The second law also provides us with a formal definition for **mass**: mass is the source of inertia; it is that property by which an object resists a change in its state of motion. The greater the mass of an object, the larger the force must be to produce a given acceleration. A change in speed or direction, or both, is an acceleration and requires a net force. Thus an acceleration can occur if just the direction of motion, and not the speed, changes. When you drive around a curve, you feel yourself pushed toward the side of the vehicle, even if the needle of your speedometer never moves. Moreover, an acceleration can be positive or negative; either the speeding up, or the slowing down, of a body is an acceleration. A negative acceleration, that is, the slowing down of a body, is often called a deceleration, but in physics the word acceleration covers both cases.

Newton's second law in terms of linear momentum

The second law can also be written in terms of *linear momentum*. In Newtonian mechanics, the linear momentum of a body is simply its mass times its velocity, that is, $p = mv$, where p is the symbol generally used in physics for linear momentum. From the definition of acceleration, it follows that the change in momentum with time is just ma . But this immediately tells us that the second law is equivalent to the statement that a force is *that quantity which causes a change in the linear momentum* of a body. The expression of the second law as a change in momentum is more general than its formulation involving acceleration, since a change in momentum can occur because of a change of mass as well as a change in velocity. Of course, the mass of an isolated object never changes in Newtonian physics, but the concept of momentum enables systems to be treated in which mass can change. A favorite textbook example of such a system is an initially empty boxcar rolling under a hopper while being loaded with coal. It is possible to compute the force on the boxcar much more easily by the application of the momentum law, than by attempting to calculate the accelerations of all parts of the system involved.

Since a force causes a change in momentum, it follows that if no force acts, the momentum of a system does not change. Newton's first law, or the law of inertia, is thus generalized to the law of **conservation of momentum**, which states that *the linear momentum of a system never changes as long as no external force acts*. The law of conservation of momentum is considerably more powerful than the law of inertia alone, since it permits such complicated systems and interactions as collisions, compound objects, and so forth to be handled elegantly by relatively simple mechanics. More importantly, the conservation of linear momen-

tum is more fundamental to the laws of physics than is the bare law of inertia. Unlike the force-acceleration forms of Newton's laws, the momentum laws can be readily extended to more advanced physics, such as special and general relativity and quantum mechanics. Even deeper, the law of conservation of momentum can be shown to arise from fundamental symmetries of space. Momentum is one of the basic quantities of the physical universe.

Force, acceleration, and velocity are all **vectors**; that is, both the magnitude (size) and direction are important. There is a special, essentially geometrical, way to add vectors that we will not treat here. It is sufficient to realize that vertical forces cause only vertical motions, while horizontal forces create horizontal motions. A force that is neither strictly vertical nor strictly horizontal can be broken into *components* along those directions; its vertical component can be added to any other vertical forces, and similarly for the horizontal component. As an example of the vector nature of forces, consider a cannonball shot straight out from a level cannon. A force is required to start the cannonball into horizontal motion, by Newton's first law; that force is supplied by the explosion due to the gunpowder, and the cannonball is then accelerated in obedience to the second law. After the cannonball exits the barrel, there is no further horizontal force upon it. Therefore, again by Newton's first law, the cannonball should continue to move at a constant horizontal speed in a straight line. But if the cannon is fired in a gravitational field, there is always a vertical force upon it. How does that affect its motion? The vertical force of gravity cannot influence the horizontal motion of the cannonball, but it does affect the vertical component; it causes the cannonball to fall to Earth. The combination of straight-line falling to the ground, with acceleration, and straight-line motion horizontally, with constant speed, creates the net curved motion of the cannonball, which is a mathematical curve called a *parabola*. The rate of fall of the cannonball is *exactly* what it would be if it had been simply dropped from its initial height. That is, if one cannonball were dropped at the same instant and from the same height as a second ball was fired from a horizontal cannon, both balls would hit the ground at exactly the same time! The horizontal distance traveled during this time interval by the second ball would depend upon its muzzle speed, of course; this effect accounts for the difficulty in observing the fall of a fast projectile such as a bullet, since it travels a great distance and generally strikes a target before it has time to fall far.

If, rather than pointing the barrel horizontally, the cannon were aimed upward, the projectile would gain a little more time because the time of travel would now be that interval required for it to rise to some maximum height and then fall back to Earth. But by firing upward it may lose some horizontal distance, because the ball's initial velocity now has a vertical component, which does not contribute to crossing the horizontal distance to the castle under bombardment. These two competing effects must be balanced to produce the maximum range. If complicating factors such

A vector consists of a magnitude and a direction, and can be represented as an arrow

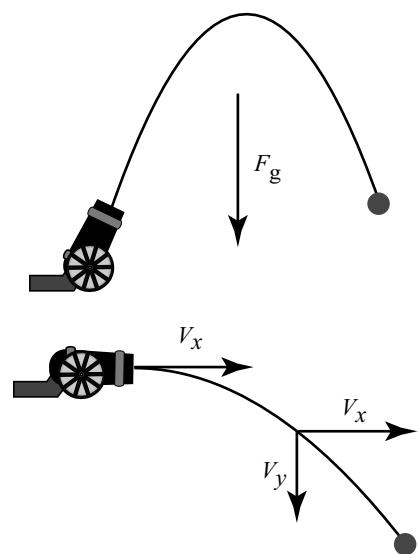
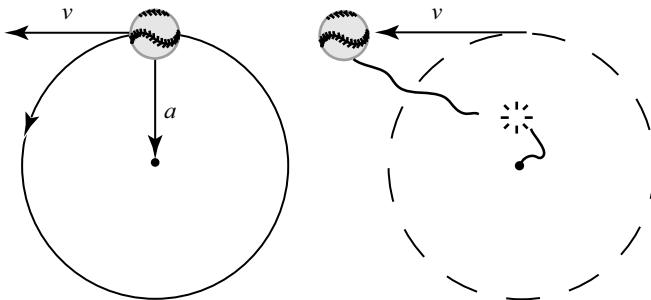


Fig. 3.2 Trajectories of projectiles in a constant gravitational field. The horizontal component of the velocity, v_x , is unaffected by the force of gravity, F_g , which acts only in the vertical direction.

Fig. 3.3 A ball tied to a string and whirled about a central point (left) moves in a circular path due to the force exerted by the string. The force, and hence the acceleration a , is directed from the ball toward the center. This is known as a centripetal force. If the string breaks (right), there is no force upon the ball and it moves in a straight line with constant speed.



as air resistance can be ignored, it can be shown that an initial angle of 45° is optimal for range.

Circular motion is not natural motion; it is forced, accelerated motion

Now we can understand why circular motion is not a natural motion. It is, in fact, accelerated motion. It has a uniform speed, but the direction of motion changes constantly. Without the force of gravity, planets would not orbit but would travel forever through space in a straight line. Gravity causes them to bend constantly, deviating from the straight line they would otherwise follow. We can also now see how to correct the misconception many people hold about circular motions. Suppose you attach a ball to a string and whirl it around your head. The tug you feel from the string tells you that a force exists. The force is exerted by your hand, and is transmitted through the string to the ball. What if you did not tie the string securely, and the ball slips away? What will be its subsequent motion? Once it is freed of the force from the string, then, according to Newton's first law, it will fly off in a straight line, *not* continue its circular motion. This is an easy experiment to try. (Simply let go of the string in order to remove the force.) Careful observation, ignoring any preconceptions you might have, will show that the ball does, indeed, move away in a straight line.

We now know how to create changes in a body's motion. Push it, pull it, or exert some other kind of force on it, and it accelerates. But if you act on something, such as push against a stalled automobile with its transmission in neutral, are you yourself unaffected? Are you able to exert forces on objects without any back reaction? Obviously this is not the case; applying forces to physical objects has consequences for you. If you push on something, it pushes back on you. The exact relation is one of equality, leading to:

Newton's third law of motion: *For every action, there is an equal and opposite reaction.*

Newton's third law

This law is easy to misunderstand, and probably causes more confusion than the other two put together. The action and reaction forces always act on *different* bodies. Body A exerts a force on Body B; and Body B exerts an equal and opposite force on Body A. Misunderstanding of this law might lead one to wonder how a horse can pull a wagon. The horse exerts a forward force on the wagon, but the wagon exerts an equal backwards force on the horse. How can they move? It is true that the

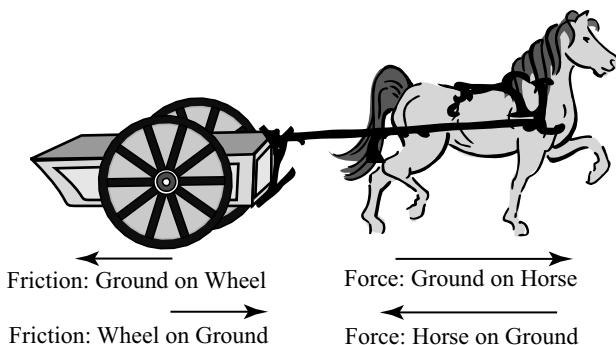


Fig. 3.4 Forces on a wagon being pulled by a horse. The net force is the (vector) sum of all forces. In this case, the net force on the horse and cart is the difference between the pull of the horse and the friction of the ground. The direction of frictional force is always opposite to the direction of the motion. If friction and pull balance exactly, the horse and cart will move at a constant velocity; if there is an excess, they will slow down or speed up. If the horse exerts a force in a direction different from the current direction of motion, the cart will turn.

wagon pulls backwards on the horse. If you have ever pulled a wagon, you are aware of the stretching of your arm, caused by the backwards pull of the wagon on you. But this is not the correct question. The horse's hooves push against the ground, and it is the reaction of the *ground* upon the *horse* that ultimately moves the wagon.

A familiar example of Newton's third law is the kick of a gun or cannon. Not everyone has ever fired a gun, but those who have, have experienced this phenomenon first hand. The explosion of the powder within the gun exerts a considerable force on the bullet. By Newton's third law, there is an equal and oppositely directed force upon the gun. This force causes the gun to accelerate, in the opposite direction from the acceleration of the bullet. Hence a force must be exerted to bring the gun to rest after its recoil. The object against which the gun is braced, often the shooter's shoulder, produces a force against the gun that, again by Newton's third law, exerts an equal and opposite force against the shooter, producing significant effects such as a bruise. The amplitude of this force depends upon the acceleration (deceleration, if you will) of the gun. The more slowly the gun is brought to a halt, the smaller the deceleration and hence the less will be the force; conversely, a rapid deceleration requires a large force. Padding on the shoulder of a shooting coat helps to slow the deceleration, and thus to reduce the force.

Does Newton's third law also imply that if the Earth attracts a brick, then the brick attracts the Earth? Indeed it does. Why, then, does the brick fall to the Earth, and the Earth not rise to the brick? The answer is found in Newton's second law. The mass of the Earth is so much larger than that of the brick that, although in principle the Earth does move, in practice, its acceleration due to the brick is unmeasurably small. The brick, on the other hand, acquires from the same magnitude of force a very large acceleration and crashes to the ground, to the hazard of anything in its path.

Examples of reaction forces

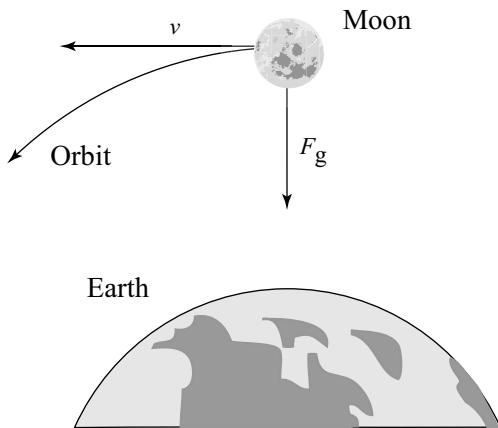


Fig. 3.5 The direction of the gravitational force F_g acting on the Moon is toward Earth. The Moon's instantaneous velocity v is perpendicular to the line joining its center to the Earth's center, but the constant centripetal gravitational acceleration produces a circular orbital path.

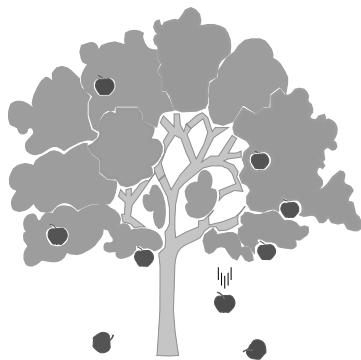


Fig. 3.6 Newton's insight was that the gravity felt on the Earth that caused the apple to fall, extends out to the Moon and accounts for its motion.

The law of universal gravitation

Now that we understand the laws of motion, we may see how they apply to gravity. According to his reminiscences, the basic ideas came to Newton when he was home in Lincolnshire during the plague in 1665. Many decades later, a younger friend reported that the aged Newton told him over tea how his thoughts turned to gravity when, upon watching an apple fall, he began to contemplate that the same force that caused the fall of the apple might also account for the orbit of the Moon. (Newton did not mention being hit upon the head by the apple, and that detail is probably just a bit of legend embroidered upon this account, if there is any truth to the story at all.) It was perfectly well understood at the time that there was some force that causes objects to fall to the ground, a force called gravity. Newton's bold leap was to imagine that the force extended not only to the surface of the Earth, but to the distant Moon.

If the Moon were moving according to Newton's first law, it should travel in a straight line. Since its path is curved its velocity changes, and hence there must be a force causing this acceleration. The force must be directed toward the Earth, or, more precisely, along the line joining the center of the Moon to the center of the Earth. Newton calculated the acceleration required to keep the Moon in orbit and found it to be about $1/3600$ as great as the acceleration due to gravity at the surface of the Earth, a quantity that had been measured by Galileo. Newton knew, from Kepler's third law, that the force had to decrease as the distance between the bodies increases. He conjectured that the force varied as the *inverse square* of the distance. Since the distance to the Moon is close to 60 times the radius of the Earth, the inverse square law is consistent with the observed acceleration. Newton later wrote that he "...thereby compared the force requisite to keep the Moon in her Orb with the force of gravity at the surface of the earth, and found them to answer pretty nearly." Thus he was able to conclude that gravity is, in fact, described by an inverse square law.

There was, however, one stumbling block. It is not obvious what the distance between the Earth and the Moon should be. Newton had assumed that the Earth, an extended body, attracts the Moon as if its mass were concentrated at a point at the center. This seems to be a reasonable approximation for the Earth–Moon system, but what about the Earth–apple system? Yet Newton’s estimate indicated that even for the apple, the Earth attracted it as if all its mass were concentrated at the center. In order to prove why this should be the case, Newton was forced to invent a new system of mathematics, integral calculus.² With integral calculus in hand, Newton was able to prove that the gravity of a spherically symmetric body is the same as that produced by the equivalent amount of matter concentrated in a point at the body’s center.

After determining the general form of the gravitational force law, Newton was obliged to evaluate the explicit formula; ratios alone would not be adequate for calculations. Newton was aware of Galileo’s demonstration that all masses fall with the same acceleration in the gravity of the Earth. He even repeated and improved upon Galileo’s work, by using pendula whose bobs were of different masses. Newton was also able to take advantage of advances in the technology of timekeeping, in order to time the periods of oscillation of the pendula. Newton found no difference in the period for a wide variety of bobs, which confirmed the results of Galileo’s original experiments with masses rolling on inclined planes. The only way in which the acceleration due to gravity could be independent of the mass of the falling object would be if the force of gravity itself were proportional to the mass of the object. Let us write Newton’s second law with a subscript to indicate that we are referring specifically to the force of gravity; the gravitational acceleration shall be denoted with the conventional lower-case g :

$$F_g = mg. \quad (3.2)$$

But g is constant at the surface of the Earth, as determined from experiment; hence F_g/m must also be constant. Therefore, the formula for F_g must contain m , the mass of the falling body.

The next step employs Newton’s third law. The mutuality of force, as required by the third law of motion, requires that the gravitational force also be proportional to the mass of the attracting body, which we shall symbolize M . Both masses must be involved in a symmetric way, if the force of Mass A on Mass B is to be equal in magnitude to the force of Mass B on Mass A. Thus Newton arrived at the conclusion that gravity was proportional to the product of the masses, divided by the square of the distance separating them. This is known as Newton’s **law of universal gravitation**, and it can be written mathematically as

$$F_g = \frac{GMm}{R^2}, \quad (3.3)$$

All objects experience the same acceleration when falling in the Earth’s gravitational field

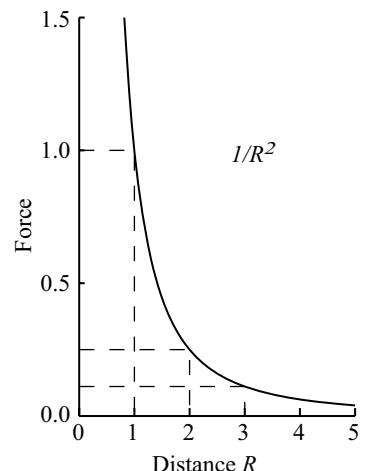


Fig. 3.7 The inverse square function. The value of the function decreases very rapidly as the distance increases.

²Gottfried Leibniz simultaneously and independently invented the concepts in Germany, and for years a nationalistic dispute raged between England and Germany over the credit for this important work. Today both men are acknowledged as the developers of calculus.

Newton's law of gravity is a mathematical description of the gravitational force between two bodies

where M is the mass of one body, m is the mass of the other, and the quantity R is the distance from the center of one object to the center of the other. The symbol G stands for the **gravitational constant**. Newton indicated G symbolically because he could not compute its numerical value; that had to be determined from experiment. In fact, measurement of G is so difficult that Newton was long dead before its value was found. Henry Cavendish, working in the last decade of the 18th century, invented a very sensitive balance, with which he was able to measure the extremely weak gravitational attraction between two spheres of known mass. Even today, however, the value of G is the least accurately known of all the fundamental constants of nature, perhaps an appropriate situation in light of the fact that gravity remains the least understood of the fundamental forces.

We can now compute the acceleration due to gravity at the surface of a planet or other spherical body by combining Newton's second law, in the form of equation (3.2), with his law of universal gravitation, equation (3.3). We obtain

$$g = \frac{GM}{R^2}, \quad (3.4)$$

The constant G must be obtained experimentally

where M is the mass of the object and R is its radius. Once Cavendish had obtained a measurement of G it then became possible to use equation (3.4) to compute the mass of the Earth, or for that matter the Sun, or indeed any other body whose gravitational acceleration and radius can be determined by one means or another. We can also now understand the meaning of the **weight** of an object. Weight is simply the force of gravity upon a given object. Using the Earth as an example, with M_E the mass of the Earth and R_E its radius, the weight of an object of mass m is given by

$$W = mg = \frac{GM_E m}{R_E^2}. \quad (3.5)$$

Weight and mass are different concepts

At the Earth's surface the radius R_E is very nearly constant, and so the force of gravity seems to be the same everywhere. (The Earth is not quite a perfect sphere, and its rotation introduces additional effects, but the corrections are small.)

Equation (3.5) shows that the weight of an object varies with distance from the center of the Earth. We do not ordinarily notice this effect because we hardly ever travel far enough from the Earth's surface for g to be perceptibly different. At the greatest height to which most of us will ever travel, the 30,000 feet (9144 m) of a cruising jetliner, the acceleration due to gravity still has 99.7% of its value at the surface of the Earth. Thus for most practical purposes, g is a constant. However, delicate instruments called gravimeters can measure tiny changes in g . They can even detect the effects due to a local mass concentration in the crust of the Earth, such as a nearby hill or perhaps an accumulation of a relatively massive mineral or ore. Weight also varies with the mass and size of the planet. Greater mass, smaller radius, or both, increase the gravitational acceleration. On the surface of the Moon, both the mass

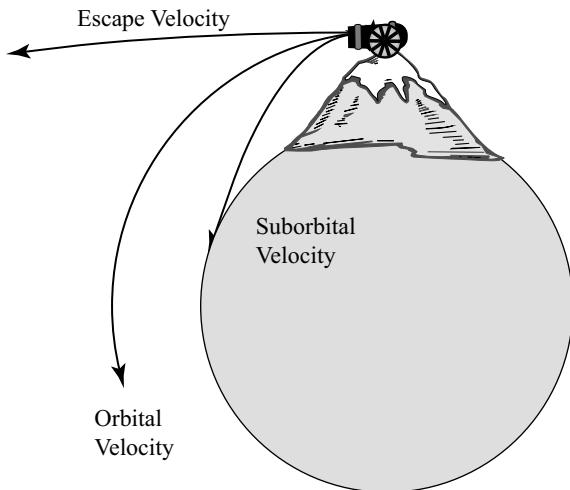


Fig. 3.8 Some possible orbits for a cannonball shot in a direction parallel to the surface of the Earth from the top of an extremely high mountain. If the initial velocity is too low, the object falls back to Earth. At a velocity of about 7.8 km s^{-1} , the Earth curves away from the cannonball at a rate that exactly matches its speed, resulting in a circular orbit. At about 11 km s^{-1} the cannonball would leave Earth's gravitational field forever.

and the radius of the planet are much smaller than they are on Earth. The Moon has only 1.2% of the mass of the Earth and its radius is just 27% that of the Earth, but the effect of the reduced mass dominates that of the smaller radius; the value of g on the surface of the Moon is about one-sixth of its value on Earth. Conversely, Jupiter is so much more massive than the Earth that the gravitational acceleration at the top of its cloud layers is 2.5 times as great as it is at the surface of the Earth.

The MKS³ unit of force is called, appropriately, the *newton*. Since weight is a force, the correct MKS unit of weight is also the newton. The everyday use of the kilogram, which is the MKS unit of mass, for weight is a convention. (Nearly all scales in common use measure force, not mass *per se*.) But, in general, mass is our concern; you could lose weight by moving to Nepal, but that would not reduce the circumference of your waist.) In the very nearly constant gravity at the Earth's surface, the distinction is not of much practical importance, although, of course, one should keep the conceptual difference clear. On the other hand, the British unit *pound* is a unit of force, and so is correctly used for weight. The British engineering unit for mass is not very well known, except perhaps to fans of crossword puzzles; it is called a *slug*. In MKS units, the acceleration due to gravity at the surface of the Earth is 9.8 meters per second per second, or 9.8 m s^{-2} . That is, if an object falls from rest, and air resistance can be neglected, at the end of one second it will be traveling 9.8 meters per second; at the end of another second it will attain a speed of 19.6 meters per second; and so forth, until it hits the ground or air resistance balances the force due to gravity.

Once Newton had determined that gravity followed an inverse square force law, he was able to prove that Kepler's first and second laws followed necessarily. The proofs are simple with the aid of fairly elementary

MKS refers to meter, kilogram, and second, the basic units of measurement

³See Appendix B for definitions of systems of units.

modern mathematics, though Newton himself had to invent calculus almost as he went along. While the derivations are beyond the scope of our discussion here, we can comment on a few of the results. Consider two bodies in orbit around each other, one with mass M_1 and the other with mass M_2 . It turns out that Kepler's second law merely requires that the force of gravity must act only along the line connecting the centers of the two bodies; such a force is called a *central force*. Kepler's first law narrows the possibilities; it requires either that the force obey the inverse square law, or else that it must increase linearly with distance. As Newton realized, of course, gravity must decrease with distance; hence Kepler's first law, combined with this observational fact, pins down the form of the force law to an inverse square. Having confirmed that the inverse square law was correct, Newton then was able to derive Kepler's third law. Since the full formulation for the gravitational force was now available, Newton was also able to work out the correct mathematical expression for the third law. The exact formulation turned out to involve the *sum* of the masses of the two orbiting bodies, as well as the distance separating them.⁴ In the solar system, the mass of the Sun is completely dominant, and so the sum is essentially equal to the mass of the Sun alone. As a consequence, the relationship $P_{\text{yrs}}^2 = R_{\text{AU}}^3$ holds for all the planets, for all practical purposes; if this had not been true, Kepler might never have discovered this law in the first place. In contrast, if we were to study a binary star system, the masses of the two stars might well be comparable. For such a system we could measure the period of their mutual orbit, and if we could resolve them telescopically we might be able to determine the size of the orbit, but without additional information we could at best find the sum of the two masses, not each individual mass. Fortunately, in cosmology we generally want to know the *total* mass of large systems, and do not need to determine the masses of specific components of the systems. Kepler's third law, as modified by Newton, thus enables us to measure the masses not only of stars, but of galaxies and clusters of galaxies. Kepler's third law thus provides a means to weigh the universe.

The three laws of motion and the law of gravitation are the fundamental relationships that make up Newtonian dynamics. At last, after more than seventeen centuries of fumbling, humanity could comprehend the motion of the heavens. The Sun is the center of attraction of the solar system; the planets and comets orbit it. The motions may be described by a succinct, precise set of laws. The impact of this discovery upon European cosmology cannot be understated. Just as the rediscovery of Aristotle's writings paved the way for the Renaissance, the elucidation of the laws of motion was a factor in the shift in thought known today as the Enlightenment, and for the industrial revolution that accompanied it.

⁴Newton showed that the precise formula for Kepler's third law is $G(M_1 + M_2)P^2 = 4\pi^2 r^3$, where r is the length of the semimajor axis of the ellipse describing the orbit and P is the period of the orbit. In this equation we may use any set of consistent units, and are not restricted to years and astronomical units.



Fig. 3.9 Edmund Halley (1656–1742), the English astronomer who persuaded Newton to publish the *Principia*. His application of Newton's laws to cometary orbits enabled him to predict the return in 1758 of the comet that now bears his name. (Courtesy of Yerkes Observatory.)

Once these laws were disseminated among the intellectual elite of Europe, humanity's understanding of the cosmos increased rapidly. The new mechanics were quickly applied to many observations and experiments. For example, Newton's friend Halley conjectured that the bright comets observed in the years 1531, 1607, and 1682, all of which shared some similarities, might actually be the same object. Halley worked out an orbit for what is now called Halley's Comet, and predicted it would reappear in 1758. He did not live to see his prediction validated. The return of Halley's Comet was first spotted by an amateur astronomer on Christmas night of 1758. It has returned faithfully, approximately every 76 years, ever since. It is fitting that the man who played such a role in the publication of Newton's masterwork should be immortalized by his own application of Newton's laws.

Newton's laws are extraordinarily simple in form, but unfortunately it is difficult to compute the consequences of the laws for gravitating systems of more than two objects. Mathematicians have shown that it is impossible to find *exact* analytic solutions for the mutual orbits of three or more bodies. However, approximate solutions may be found with pencil and paper, provided that one body is much more massive than the others. In that case, orbits are determined, in the main, by the two-body equations, with small corrections. A planet's orbit is very nearly determined by considering just the planet and the Sun. The gravitational influence of other planets produces small perturbations on

Newton's laws are predictive

that orbit. Newton had known only the five planets familiar to the ancients, but he suspected that the mutual attraction of Jupiter and Saturn might be detectable, and he even asked John Flamsteed, the astronomer royal, whether the two showed any anomalies in their orbits; none were observed at the time.

In 1781, more than half a century after the death of Newton, William Herschel discovered Uranus by direct observation, using a telescope he designed and built himself. Uranus is, in principle, visible to the unaided eye, but barely so; Herschel's knowledge of the sky enabled him to spot a dim star where no star should have been. Astronomers duly recorded observations of Uranus and computed its fundamental orbit. By 1845, the data were sufficiently precise to show that the orbit of Uranus could not be explained by perturbations from the known planets. This small anomaly soon led to a great triumph of Newtonian mechanics: the prediction of an unseen planet. John Adams in England and Urbain Leverrier in France simultaneously predicted that a new planet must lie beyond Uranus, and both gave a location for that planet. At first the observers paid insufficient heed to these predictions, especially in England; Adams' work was ignored by George Airy, the Astronomer Royal at the time. Finally, in 1846 an astronomer at the Berlin observatory discovered the new planet on his first attempt, within 1° of its predicted location. In keeping with the practice of naming planets for Graeco-Roman deities, the new planet was christened Neptune, for the god of the sea.

Compared with what was possible with the Ptolemaic tables, the power of Newton's mechanics was intoxicating. The whole universe, for all time and space, seemed within the grasp of humanity's understanding. The Newtonian universe was infinite in extent and populated evenly with stars similar to the Sun. Each star had its own mass and a specific instantaneous velocity. Given the mass of every planet and star in the universe, and their velocities and positions at one instant in time, Newton's equations are fully deterministic, predicting both the future and the past evolution. The gravitational law provides a force, the second law determines the acceleration, the acceleration determines the velocity, and the velocity determines the new position. The practical difficulties of actually computing the evolution of the universe are not so important. The watershed was the transformation of the universe from something intrinsically mysterious and unknowable, to something deterministic and calculable.

In modern times the availability of tremendous computing power reduces somewhat the practical problem inherent in Newton's laws. The equations of gravitation among large numbers of bodies are routinely solved with great accuracy on computers; one example of an application to cosmological research of these N -body simulations is their use in the investigation of the formation of galaxies in the universe. However, it turns out to be impossible to predict orbits for all times with arbitrary accuracy. Self-gravitating systems are known to be *chaotic*; eventually, very small errors in our knowledge of the current orbits become large

Newton's laws predicted the existence of Neptune

The apparent simplicity of Newton's laws masks the complexity of many-body systems

errors in our projections of future orbits. For the solar system, with relatively few bodies and one very dominant mass, these errors grow only over billions of simulated years, so for the needs of determining the orbits of, for example, space probes, we may solve Newton's equations to any desired precision. Nevertheless, the chaotic behavior of gravity shows that we can still find surprises in Newtonian mechanics.

Newton was well aware of the majesty of his accomplishments, yet he was also aware of their limitations. A particular difficulty, for which others criticized him, was the appearance that gravity exerted its influence instantaneously at a distance, with neither an intermediary nor obvious causal contact. Newton conceded that he could not find the cause of gravity but it was, for the moment, enough to elucidate its effects. For insights into cause, the world would have to wait for Einstein.

I do not know what I may appear to the world; but to myself
 I seem to have been only like a boy playing on the seashore,
 and diverting myself in now and then finding a smoother peb-
 ble or a prettier shell than ordinary, whilst the great ocean
 of truth lay all undiscovered before me. (Isaac Newton)

The age of the Earth

Just as Aristotle's cosmology fitted the attitudes of the Middle Ages, so did Newton's cosmology suit the prevailing philosophy of the Enlightenment. The universe was like a grand clockwork, the stars and planets turning to the pull of gravity like the bearings of a finely balanced watch. It was a confident age, when knowledge of both science and technology increased rapidly. The Industrial Revolution was stirring, and Europe was well established in its colonial adventurism. Among the educated, affluent classes of both Europe and North America, a popular theology was deism, which was heavily influenced both by Newtonian cosmology and by the growing precision in technology. Deism views the universe as a kind of majestic machine, created by a master machinist and set into eternal motion; it is natural law that reveals the divine. Newton's clockwork universe ticked along for almost two centuries before new complications arose, which once again changed our views of the universe.

Newton's clockwork universe

The geologists

If cosmology was the grandest science, based upon the stars, geology began as one of the humblest, a purely practical exercise in locating exploitable minerals, planning roadbeds and canals, and the like. Perhaps the old Aristotelian notion of the Earth as debased persisted; in any case, the study of the Earth for its own sake did not begin to become established as a science until the 18th century. However, just as the contemplation of the heavens above produced an understanding of the vastness of space, consideration of the rocks at our feet was to produce an awareness of the vastness of time.

Geology hints at the antiquity of the Earth

The concept of uniformitarianism

Geologists noticed very early on that many rock formations were stratified, apparently built layer upon layer by some process. In 1669 Nicolaus Steno published the suggestion that older rocks were below and newer ones were on top, an idea that seems perfectly obvious today; but at the time, all rocks were thought to be the same age. It slowly became clear that rocks were marked by the history of the Earth, and in the late 1700s James Hutton proposed the theory of *uniformitarianism*, the assertion that the same geological processes that we observe today, such as wind, water, and volcanism, also operated in the past. Geologically, the past can be explained by an understanding of the present. Uniformitarianism did not gain immediate favor, as it conflicted with the prevailing beliefs in Europe at the time. Nevertheless, the very thickness of the layers of rock could not be ignored, and in the early 19th century geologists began to entertain the idea that the Earth might possibly be very old. Within some of the layers fossils were found, strange traces of creatures that matched no known living animals. At first many Europeans believed that these animals were still alive somewhere else; but as more and more of the world became known in Europe, this belief became increasingly untenable.

The theory that found favor during much of the early 1800s, especially in France, was *catastrophism*, the belief that the Earth had experienced numerous and frequent upheavals in the past, each catastrophe wiping out the animals of that geological layer. The extinct animals were replaced, either by a separate creation or else by colonization by animals from other regions. But there were always doubts. Whatever the explanation for these mysterious imprints, the geologist William Smith demonstrated during the end of the 18th century and the beginning of the 19th that a given type of rock layer was uniquely associated with a particular set of fossils. The strata could be arranged in relative order by examination of the fossils they contained. This was not inconsistent with catastrophism, but the depth of the layers implied such a large number of catastrophes as to be uncomfortable. Moreover, there were resemblances in animals from one layer to another. By the 1830s geologists had realized that the strata showed a progression of complexity of the fossils they bore. The oldest rocks contained no detectable fossils. Next came layers that held only invertebrates, and finally came newer layers which successively were dominated by fishes, reptiles, and finally mammals and birds.

One of the earliest proponents of a transformation theory in biology was the French scientist Jean-Baptiste Lamarck. Lamarck held that species changed over time, gradually and in response to environmental conditions, a view which put him into direct conflict with catastrophism. Lamarck's explanation for such changes was that acquired traits could be passed from parent to offspring; the most famous example is the giraffe, stretching its neck to reach higher leaves and passing the elongated neck to its offspring. We know now that acquired traits cannot be inherited, but during the whole of the 19th century, the mechanisms of inheritance were entirely unknown. Gregor Mendel's work in establishing the dis-

crete and predictable nature of inheritance was carried out from 1862 to 1865, but it was ignored for four decades. Lamarck, like many of his time, believed in Aristotle's *Scala Naturae*, the great Ladder of Life, a hierarchical arrangement of creatures in order of increasing perfection with the pinnacle of life, humans, at the top. The driving force for evolutionary change, in Lamarck's view, was not survival, but an urge to climb the ladder toward greater complexity. Although Lamarck's theories have been completely discredited, it should be noted that it was the first consistent proposal that animals change over time, that species are not fixed and perfectly suited for their niches.

Meanwhile, in Great Britain, Charles Lyell published between 1830 and 1833 his *Principles of Geology*, a book that placed uniformitarianism on a firm foundation. Lyell's work is often considered the beginning of modern geology, and it clearly showed that the Earth was ancient, although no one knew at the time how old.

The Aristotelian concept of a natural hierarchy of life

Darwin and Wallace

Charles Darwin was the scion of a wealthy and influential family. His paternal grandfather Erasmus was a major figure in the elite circles of the day, and his maternal grandfather was Josiah Wedgwood, founder of the famous china and pottery company. Charles, however, was a woolgatherer, indifferent toward his studies and most enthusiastic about wandering the countryside, collecting specimens of interesting animals and plants. When Charles received an offer in 1831 from Captain Robert Fitz-Roy to travel as naturalist aboard the HMS *Beagle* on a five-year voyage of exploration, he eagerly accepted. Darwin's adventures were to write the final chapter of the Copernican revolution, by removing humankind from its assumed splendor as a special kind of creature.

One of the books that Darwin took along with him was Lyell's first volume on geology. As he observed both the variety and the similarities of animals all around the world, Darwin came to see the evolution of species as itself a form of uniformitarianism; the same processes occurred throughout the history of the Earth, leading to slow and gradual changes in the animals who occupied it. It was the competition for survival, and the survival of offspring, that drove these changes. In any generation, those best suited to their environments left more offspring, of which in turn the best adapted reproduced most successfully. The natural variations in individuals were the raw material of change. Some of the offspring of the giraffe's ancestors had necks a little longer than others; by their ability to reach higher leaves they gained a better diet and produced more offspring, of whom the longest-necked survived best. There was still no understanding of the biology of heredity, but Darwin was thoroughly familiar with artificial selection, especially pigeon breeding, by which the breeder chooses for breeding stock those young that display some desired characteristics. Of course, nature could work in a similar manner. Natural selection, operating upon the inherent variability of a population, could, over the eons of time provided by the new

Darwin's extensive observations lead through inductive reasoning to the principles of evolution

understanding of geology, produce the great array of species from a few ancestors.

By the time he returned from his journey in 1836, his ideas had already crystallized, but Darwin had little courage to confront their implications. He feared dissent and the disapproval of his family. He thus delayed publication for nearly twenty years, until a fateful letter arrived from a young man who, like Darwin many years before, had set off to see the world. The young man was Alfred Russel Wallace. Unlike Darwin, Wallace had grown up and lived in poverty and hardship. Wallace had also traveled the tropics, and he had reached the identical conclusions as Darwin. In fact, it was Wallace who first broached the topic to the wider world while Darwin dallied and procrastinated, endlessly reworking his notes. Wallace published a small paper in 1855, proposing that species came into existence from earlier species. It was only then that Darwin was persuaded to publish. Darwin's friends made arrangements for papers by both men to be read at the same scientific meeting in 1858. (Wallace, the working-class outsider, seems to have been deeply grateful at the opportunity to be heard by the learned men of British science, a profession generally reserved for members of the upper class at the time.) The credit for the theory of natural selection rightly belongs to both Darwin and Wallace, but it was Darwin who published, in 1859, the landmark book *The Origin of Species*, and Darwin whose name became associated in popular parlance with evolution.

The *Origin of Species* was described by Darwin himself as "one long argument." The book marshals many facts and shows how simply they fit the hypothesis of natural selection, but at the time no mechanism was known that could account for the process of gradual change in species. It was not until the beginning of the 20th century that the work carried out by Gregor Mendel in his monastery garden was rediscovered and replicated, leading to an understanding of heredity. The elucidation of the biochemistry of inheritance began only in 1944, when it was established that the unit of heredity, the gene, was composed of the molecule deoxyribonucleic acid, or DNA. Though much remains to be learned about genetics, the broad outlines are now well understood. Moreover, evolution itself, in response to environmental pressures, has been directly observed on small scales. The examples with the most ramifications for human societies are the development of drug resistance in bacteria and pesticide resistance in many insects. A dose of antibiotics may kill most, but not all, of a population of bacteria. Those that survive carry a trait that enables them to resist destruction by the drug. The resistant bacteria are able to reproduce extravagantly in the ecological space cleared by the deaths of their erstwhile competitors, and the frequency of the gene that endows the bacteria with resistance increases. Now, after fifty years of routine treatment with antibiotics, many common bacteria are resistant to drugs such as penicillin, and pharmaceutical companies must engage in a constant search for new substances that are effective, at least until new resistant strains arise.

The elucidation of genetics reveals the mechanism of evolution

If all creatures gradually changed through a process of natural selection, then the Earth could not be young. The process of evolution requires a great deal of time. Yet the geologists had already reached this conclusion; their evidence pointed to a slowly evolving Earth, and antiquity of the Earth provided time over which biological evolution could occur. The obvious changes that had taken place in the crust of the Earth itself made the idea of change of species tenable. The controversy made an accurate estimate of the age of the Earth one of the most important scientific problems of the late 19th and early 20th centuries.

Biological evolution requires time

Ironically, one of the first challenges to face the new theories came from physics. Lord Kelvin (William Thomson) and Hermann Helmholtz independently computed the age of the Sun to be only approximately 100 million years, a figure that contradicted the evidence from geology, and which did not provide sufficient time for evolution to occur. With perhaps the unfortunate tendency of some physicists toward arrogance, Kelvin declared that the geologists were wrong, since the Earth cannot be older than the Sun; he did, however, hedge by commenting parenthetically that he had considered only the known laws of physics, and new phenomena could alter his result. Kelvin's mathematics were correct but his assumptions were wrong; he had assumed that the Sun shone by means of the release of its gravitational energy as it contracted, in which case it would, indeed, be young. In fact the Sun, as well as all other stars, is powered by nuclear reactions, a physical process unknown in Kelvin's time. The physics of the atomic nucleus was developed during the first thirty years of the 20th century, and the notion occurred to several scientists, including Arthur Eddington and George Gamow, that nuclear reactions might play a role in stars. The fusion of protons was proposed as early as the 1920s as an energy source for the Sun, but nuclear physics was barely understood at that time, and the details were not correctly worked out. It was not until 1938, after further progress in the theory of quantum mechanics, that Hans Bethe elucidated some of the reactions by which the stars shine.

The discovery of nuclear physics not only removed the apparent age problem of the Sun, but also provided a means for directly measuring the age of the Earth. By the 1920s the technology for **radioactive dating** had become established. For the dating of rocks, one of the most useful isotopes is uranium 238 (^{238}U), which decays to lead 206 (^{206}Pb) at a known rate. By comparing the ratio of ^{238}U to ^{206}Pb , it is possible to determine the time since the rock solidified into its present form. The technique works best for igneous rocks, those brought up from the depths of the Earth's mantle by volcanic action, since such rocks contain the greatest quantities of radioactive elements; however, the principle can be applied to anything in which radioactive nuclei are present. The results indicate that the oldest rocks on Earth are approximately 3.9 billion years old. Material from meteorites, whose surfaces were never molten and are essentially unchanged from their origin, show ages of approximately 4.5 billion years. The oldest of the Moon rocks returned by the Apollo astronauts are about the same age as the meteorites.

Natural radioactivity provides a chronometer

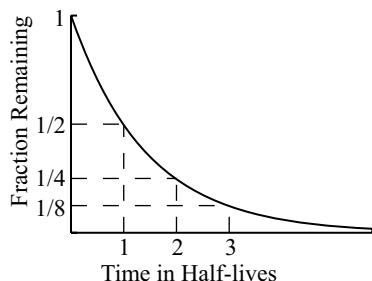


Fig. 3.10 Radioactive dating. Half of the atoms of a radioactive element in a sample will decay after each half-life. By measuring the ratio of the remaining amount of radioactive element compared to the quantity of its decay product, the age of the sample can be determined.

Various other dating techniques, as well as theoretical computations of the age of the Sun, agree quite well that the solar system is 4.6 billion years old. The most ancient established fossils, of bacteria, have been found to be 3.5 billion years old, very nearly the same age as the oldest surviving rocks.

The age of the solar system is estimated to be about 5 billion years. The Galaxy must be older than the solar system; the best estimates of its age come from determinations of the age of globular clusters, which are thought to be the oldest objects in the Milky Way, and from the ages of ancient white dwarf stars. These methods give an age for the Milky Way of more than 10 billion years. The universe itself must be older still. Indeed, the age of the universe is now calculated at just under 14 billion years. We shall discuss this in greater detail in later chapters.

Taking down the ladder

In the medieval Ptolemaic cosmology, the physical construction of the cosmos was hierarchical. At the center was Hell, the basest and lowest possible state. At the other extreme, outside the sphere of the stars, was the realm of the spirits. Humanity lay in the middle, on the surface of the Earth. Similarly, the great Ladder of Life placed humans at the peak of the Earthly species, but falling short of the perfection of heavenly beings. From Copernicus onward, the understanding has grown that the universe is not static, not perfect and immutable, but dynamic and ever-changing. The Earth, assumed through most of human history to be stationary and central, is a small chunk of rock in orbit about a middling star. What Copernicus did to the heavenly spheres, Darwin did to the Ladder of Life. Humans sit at no pinnacle, either at a physical center of the cosmos or at the peak of biological perfection. Throughout the history of the Earth its life has always been, and always will be, dominated by bacteria. Bacteria are found in every environment that can support life, and are by far the most common organisms. For more than two billion years, bacteria were the *only* life forms on Earth. Eukaryotes, cells with true nuclei, first appear in the fossil record scarcely a billion and a half years ago. Multicellular organisms have existed for only approximately 750 million years. The genus *Homo*, to which modern humans belong, arose on the plains of Africa some two million years ago, while anatomically modern humans go back at most a mere 250,000 years. In contrast, the dinosaurs were the dominant vertebrates for over 100 million years. Most of Earth's history took place without the presence of humans. If we disappeared, the Earth and its major life form, the bacteria, along with whatever other organisms might exist at the time, would continue unperturbed.

The Earth is but one small planet orbiting around one ordinary star; should it be distinguished as the only place in which life occurs? A straightforward adoption of the Copernican principle would argue that we are not alone. Life may not be common in the universe, and it may not exist anywhere else in the Milky Way. But if planets formed

around one unexceptional star by a process which, to the extent that it is understood, does not require unusual conditions, then planetary systems must be abundant, especially around stars that lack binary partners. It is true that life is fairly sensitive, placing demands upon the conditions it requires, at least for the carbon-based life with which we are familiar. Life, as we understand it, requires reasonable stability of star and planet, the presence of a good solvent such as liquid water, and protection from disruptive radiation from the star, so that the weak chemical bonds that hold together the complex molecules of life are not broken. But under the right conditions, the great antiquity of life on Earth indicates that it develops readily. Of the unknown trillions of stars in the uncounted billions of galaxies, it is difficult to argue that there cannot be other planets that support life. Whether intelligent life would exist on such planets we cannot, as yet, say. The development of intelligent life, or at least life forms that are capable of asking questions about the universe in which they live, does not even seem to have been inevitable on Earth.

For many such cosmological questions, we have no definite answers. But we have come far from the geocentric, anthropocentric world of Aristotle. With the realization of our true place in the universe, humankind has been forced to accept humility. In exchange we have found that the universe of which we are a part is far larger, grander, and more fascinating than could have been imagined even a century ago.

Chapter Summary

Isaac Newton formulated the laws of mechanics that describe most motions in the universe. Newton's first law defines inertial, or uniform, motion: a body at rest or in a state of uniform motion will remain at rest or in uniform motion unless acted upon by a net external force. Newton's second law defines mass as the connection between force and acceleration: $F = ma$. The first law can be understood in terms of the second: if the (net) applied force is zero, the acceleration is zero, meaning that the velocity does not change. Newton's third law states that for every action there is an equal and opposite reaction.

Newton's law of universal gravitation states that the gravitational force between two objects is proportional to the product of their masses divided by the square of the distance between them. The constant of proportionality is one of the fundamental constants of nature, the gravitational constant G . The gravitational constant was first measured nearly a century after Newton's death, and even today its value is known less precisely than are the values of other important physical constants.

Newton published his work, the *Philosophiae Naturalis Principia Mathematica*, in 1687; it is one of the greatest scientific treatises ever written. In addition to laying out the laws of mechanics still used today, Newton was able to derive Kepler's laws and to show that gravitational orbits would take the form of an ellipse.

After the publication of the *Principia*, understanding of the universe increased dramatically; combined with breakthroughs in technology, the new science led to the era historians call the Enlightenment. One application of Newtonian physics was computed by Edmund Halley, who worked out the orbit for the famous comet that now bears his name. Another was the prediction and subsequent discovery of Neptune. The Newtonian cosmos was a majestic and deterministic clockwork. The clockwork universe and the new understanding of natural law influenced philosophy and theology.

Nearly two hundred years passed before another major shift in cosmological thought took place. Just as Newtonian physics had made possible an understanding of the true size of the solar system, geology and biology eventu-

ally led to a new appreciation of the age of the Sun and planets. Darwin and Wallace developed the theory of biological evolution late in the 19th century, before the age of the Earth had been determined. By the 1920s, radioactive decay had been employed to measure the age of the Earth to be close to 4.5 billion years. This vast expanse of geological history allowed time over which biological evolution could occur.

The laws of physics provide the foundation for a particular cosmology. By the same token, discoveries about the

nature of the universe must be consistent with the laws of physics. The heliocentric cosmology of Copernicus, as clarified by Kepler, led to the need for a new theory of motion. Newtonian mechanics, in turn, created a new vision for the cosmos. Discoveries made toward the end of the 19th and the beginning of the 20th centuries led to the new physics of Einstein and, in turn, to the modern big bang cosmology.

Key Term Definitions

Newton's first law The law of motion which states that an object in a state of uniform motion will remain in that state unless acted upon by an external force.

uniform motion Motion at a constant velocity. The state of rest is a special case of uniform motion.

force That which produces an acceleration.

acceleration A change of velocity with respect to time.

Newton's second law The law of motion which states that the net applied force on an object produces an acceleration in proportion to the mass: $F = ma$.

mass That property of an object which causes it to resist changes in its state of motion; also, that property which generates gravitational attraction.

conservation of momentum The principle that the linear momentum of a system (in Newtonian mechanics, mass times velocity) remains the same as long as no external force acts.

vector A mathematical entity that has direction as well as magnitude. Important physical quantities represented by vectors include velocity, acceleration,

and force. A vector changes whenever either its direction or its magnitude changes.

Newton's third law The law of motion which states that if A exerts a force on B, then B will exert an equal and oppositely directed force on A. For every action, there is an equal and opposite reaction.

universal gravitation Newton's mathematical formulation of the law of attraction between two masses: $F_g = GM_1M_2/R^2$.

gravitational constant A fundamental constant of nature, G , which determines the strength of the gravitational interaction.

weight The gravitational force experienced by an object. It usually refers to the gravitational attraction due to a large object, such as a planet, upon smaller objects at or near its surface.

radioactive dating The determination of the age of a sample by the measurement of the ratio of the decay products to the precursor, for one or more radioactive isotopes. Radioactive dating is possible because each unstable isotope has a well-defined half-life.

Review Questions

- (3.1) When fully loaded with fuel, a certain aircraft has a mass 1.25 times greater than its mass when carrying minimal fuel. The acceleration of the aircraft for takeoff must be the same in both cases, since the length of the runway available, and the takeoff speed required, are both fixed. How much more thrust (force), relatively, must the engine exert to accelerate the aircraft for takeoff when it is fully loaded?
- (3.2) Airplanes, especially smaller ones, often “crab,” that is, fly at an angle relative to the desired direction of travel. What conditions might make this necessary?
- (3.3) Why is circular motion not natural? Why does the velocity of an object in circular motion change even though its speed is constant?
- (3.4) If an identical force is applied to two separate masses, object A and object B, and object B is four times as massive as object A, how will object B's resulting acceleration compare with object A's? Next, consider two objects of equal mass separated by a distance of 1 meter. They feel a mutual gravitational force. By what amount does that force change if the distance between the two objects is reduced by half?
- (3.5) You are an astronaut floating in space, while holding an object with a mass that is 1/100th of your mass. You throw this object in some direction. What happens to you? Is there a difference between how Newton's laws work in space and how they work on the Earth?
- (3.6) In a certain science fiction story written for youngsters, an accident causes an untethered astronaut to float away from his spaceship. Fortunately, he manages to return safely to the ship by making swimming motions with his arms. What is wrong with this? What is the difference between swimming in water and “swimming” in space?
- (3.7) Aristotle says “To keep your automobile moving down the highway requires a steady force, hence you must keep your foot on the accelerator pedal.” What would Newton say in rebuttal?
- (3.8) The Moon orbits the Earth. State how Aristotle and Newton each explained this phenomenon and compare their explanations.
- (3.9) Zorlo has a mass that is 1.5 times that of the Earth, and a radius 1.25 times greater. How large is the acceleration due to gravity at the surface of Zorlo compared to the acceleration at the surface of the Earth? (Hint: you do not need to know the value of the gravitational constant for this problem.)
- (3.10) Which of Kepler's laws enables modern cosmologists to compute the mass of a distant cluster of galaxies? How might such a measurement be performed?
- (3.11) What is the principle of uniformitarianism? Can it be applied to the universe as a whole as well as to the Earth?
- (3.12) When Lord Kelvin computed the age of the Sun, what critical assumption did he make? What was his result, and what did he conclude about the age of the Earth? What later discovery showed Kelvin's result to be incorrect, and why?
- (3.13) Astronomers sometimes try to estimate the number of planets on which technological beings might live. Part of the process is to estimate (1) the fraction of planets capable of supporting life on which life actually appears, (2) the fraction of those planets with life where some form of life achieves intelligence, and (3) the fraction of those planets where intelligent life develops the technological capabilities necessary to send radio signals. What do you think these fractions might be? Discuss your choices.

This page intentionally left blank

Part II

Background

This page intentionally left blank

Lighting the Worlds

4

Key Terms:

It is a capital mistake to theorize
before one has data.

— Arthur Conan Doyle, *A Scandal in Bohemia*

Humanity contemplates the cosmos from the restricted vantage point of a small planet near the edge of an out-of-the-way galaxy. How it is that we can learn anything at all about so vast a thing as the universe? The universe is more than its contents alone; the physical laws that govern the interactions among objects tell us the properties of the cosmos. From the smallest elementary particles to the largest galaxy cluster, the rules of the universe leave their mark. The Copernican principle ensures that by understanding how nature works here in our own backyard, we can uncover the workings of the cosmos itself. In only 300 years, science has made great progress in elucidating these rules, and may be close to an understanding of how the universe operates at its most fundamental level. Science builds upon what is known; since cosmology deals with the overall principles of the universe, it draws upon knowledge from many fields. This chapter provides a brief, and highly selective, overview of a few topics relevant to our later studies: the basic properties of matter, the fundamental forces of physics, and some properties of light. From there we shall begin our detailed exploration of the grandest science.

The nature of matter

Humans have long searched for the fundamental basis of matter. During the Middle Ages, Europeans generally held the Aristotelian view that Earthly matter consisted of four elements: earth, air, fire, and water. The heavenly bodies were made of the celestial ether, an ill defined, perfect, and immutable substance. The modern view of matter began to take shape when chemistry developed into a science in the 18th century, distinguishing itself from the mystical pursuit of alchemy. Antoine Lavoisier¹ and Joseph Priestley showed that it was possible to attribute chemical behaviors to certain substances into which most

- photon
- nucleon
- isotope
- nuclear reaction
- neutrino
- boson
- fermion
- exclusion principle
- antimatter
- strong interaction
- weak interaction
- electromagnetic force
- gravity
- energy
- potential energy
- kinetic energy
- thermodynamics
- heat
- temperature
- conservation of energy
- conservation of matter
- entropy
- spectrum
- interference
- Doppler effect
- blueshift
- redshift
- lightyear
- blackbody radiation
- equilibrium
- luminosity
- galaxy
- galaxy cluster
- supercluster

¹Despite his support for the French Republic, Lavoisier was guillotined during the aftermath of the French Revolution, apparently because he had once been a tax collector. The judge at his trial is said to have remarked, “The Republic has no use for savants.”

The structure of matter is described by atomic theory

chemicals could be broken down. These substances, which took many forms, were themselves chemically irreducible; they are the *elements*. By 1810, it was accepted that each element corresponded to a unique type of particle, an *atom*, a theory first developed in its modern form by John Dalton. The atom is the smallest subdivision of matter that retains fixed chemical properties. The enormous variety of chemicals is created by the *chemical bonds* between atoms. Combinations of two or more atoms are called *compounds*; the behavior of a compound is, in general, nothing like the behavior of any of the elements that make it up, but depends in a fairly complicated way on the elements present and how they are bonded.

For astronomy the most important elements are the first two in the Periodic Table, hydrogen and helium. Hydrogen makes up about 75%, and helium approximately 24%, of all the matter in the universe. The rest of the elements, while far less abundant, play an obviously important role: the Earth and everything on it, including humans, are made of these elements. One of the great successes of modern astronomy and cosmology is their explanation for the formation of the natural elements and their relative abundances. As we shall see, the cosmic abundances of the elements severely constrain the possible models of the universe. The atoms themselves can tell us something about the history of the cosmos.

In 1869, Dmitry Mendeleev and Lothar Meyer, working independently, arranged the known elements into a table according to their atomic weights. Remarkably, the elements were found to show regularities in their behavior that repeated themselves nearly uniformly along a column of the table. These regularities were so predictable that Mendeleev was able to shift at least one element, indium, whose atomic weight had been incorrectly determined. He left spaces for undiscovered elements, predicting not only their atomic weights but also their general chemical properties. When the first missing element, gallium, was discovered in 1875 it created great excitement, for it made clear that there was a unifying principle to chemistry that could soon be understood. After the development of atomic theory early in the 20th century, chemists realized that another characteristic of atoms, the *atomic number*, was the key to chemistry. When the elements are arranged in order not of atomic weight but of atomic number, the regularities along the columns of the modern *Periodic Table* are nearly exact.

Cosmic abundances

The discovery of the electron

The first discoveries of *elementary particles* soon clarified the Periodic Table. Atoms, while they do indeed represent the smallest particle of a particular element, are not indivisible. The *electron*, discovered by J. J. Thomson in 1897, was the first elementary particle found. Thomson was investigating cathode rays, charged beams propagating in evacuated glass tubes, which were of great interest in the late 19th century. The existence of electrical charge had been known since the ancients observed that amber (Greek *elektron*) could, when rubbed with fur, attract small bits of straw. In the 1700s Benjamin Franklin studied electricity and proposed that it was of two varieties, which he dubbed “positive” and

1 H																																				2 He	
3 Li	4 Be																																				
11 Na	12 Mg																																				
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr																				
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe																				
55 Cs	56 Ba	57 La	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Ti	82 Pb	83 Bi	84 Po	85 At	86 Rn																				
87 Fr	88 Ra	89 Ac																																			
Lanthanide series																																					
Actinide series																																					
58 Ca	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu																								
90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr																								

Fig. 4.1 The Periodic Table of the elements. The number above each symbol is the atomic number, the number of protons in the nucleus. The two lightest elements, hydrogen (H) and helium (He), are most abundant in the universe.

“negative.” It was also realized that like charges repel one another, while opposite charges attract, but until Thomson’s work little more than these few facts was known about the nature of charges and currents. When Thomson discovered that the ratio of the charge to the mass of the cathode rays was independent of the materials used to construct the tube or the low-pressure gas that filled it, he made what amounted to a leap of faith that he had discovered a new, subatomic particle; his insight was later confirmed by more exacting experiments. The discovery of the electron made it clear that charge could be associated with individual particles. The charge, if any, controls the electrical behavior of the particle. If a particle has a charge, it is either positive or negative; if there is no charge, the particle is neutral. *Currents*, such as those that power electrical devices, consist of charged particles in motion.

Even after the discovery of the electron, a great deal of confusion over the structure of the atom persisted for quite a long time. Atoms were known to be electrically neutral, but the atomic weight and number were not understood. The “plum pudding” model proposed by J. J. Thomson envisioned a structure with electrons embedded like raisins (plums, in certain baking contexts) in a “cake” of positive charge. This model was accepted for a while, but it never worked well. Ernest Rutherford set out to test it; in a series of experiments performed between 1909 and 1911, he shot beams of *alpha particles* (now known to be helium nuclei) at an extremely thin gold foil. Most of the alpha particles passed through the foil with only slight deflections, but there were a few that were deflected by large amounts, in some cases nearly reversing their directions. Though very few such extreme scattering events occurred, the result completely contradicted the predictions of the plum pudding model; at the time, no known model of the atom could explain it. Rutherford later said, “It was quite the most incredible event that has ever happened to me in my life. It was almost as incredible as if you had fired a 15-inch shell at a piece of tissue paper and it came back and hit you.”

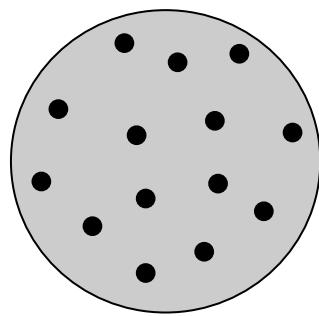


Fig. 4.2 The plum pudding model of the atom. Individual electrons are embedded in a general “cake” of positive charge.

A model for atomic structure

Fig. 4.3 Rutherford scattering by an atomic nucleus. Most alpha particles (energetic helium nuclei) shot toward gold foil are barely deflected, but those passing very close to a nucleus undergo large deflections. No such extreme deflections would be observed if the atom's positive charge were smoothly distributed; the extreme deflections provide evidence for compact nuclei within atoms.

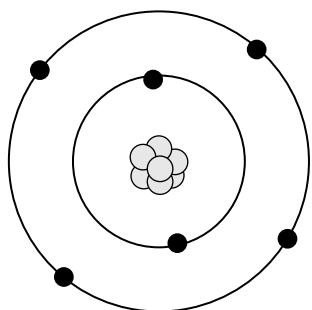
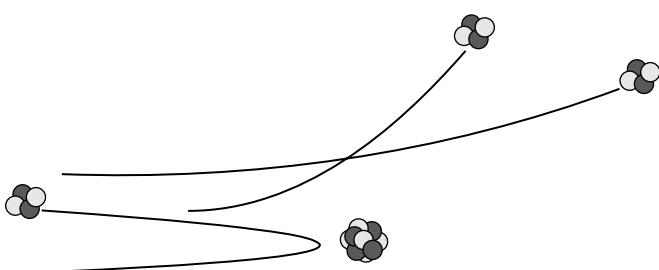


Fig. 4.4 The solar system nuclear model. Individual electrons orbit a compact, positively charged nucleus.

The Bohr atom

In 1911, Rutherford developed a new model for the atom; it consisted of a tightly packed positive *nucleus* surrounded by orbiting electrons. This theory explained his data perfectly. Most of the bombarding alpha particles, which are positively charged, passed far from the nucleus and were scarcely affected, especially since the cloud of negatively charged electrons partially cancels the positive charge from the nucleus. But a very few alpha particles happened to penetrate the electron cloud and pass very close to the nucleus, and for these particles the repulsive force was very large. Based upon his model, Rutherford was even able to predict mathematically the probability of such large deflections; the predictions fitted the data extremely well. Further work eventually led Rutherford to the realization that the atom with the smallest atomic number, hydrogen, had a nucleus consisting of a single particle. The new particle was christened the *proton*.

Almost immediately, theorists began to work out the details of this new model of the atom. The first, and most obvious, model was based on an analogy with the solar system. After all, the electrostatic force between charges obeys an inverse square law, as does gravity, and the nucleus is much more massive than the electrons, just as the Sun is far more massive than any planet. Unfortunately, the analogy broke down. By the time of Rutherford's discoveries, it was known that accelerated charged particles radiate away some of their energy in the form of light. The original solar system model of Rutherford and Arnold Sommerfeld was thus untenable; the orbiting electrons should lose energy and spiral into the nucleus, which would have had most unfortunate consequences for chemistry! Clearly, a better model for the atom was required. Niels Bohr, working on the hydrogen atom, hit upon a solution in 1913. Bohr's work was a major contribution to the nascent quantum theory, as it showed that electron orbitals were *quantized*; they could not be arbitrary. Unlike planets, electrons could occupy only discrete orbits of fixed energy. As long as the electron occupied a permitted orbital it did not radiate, but if it jumped from one orbital to another it emitted or absorbed a single *quantum*, an indivisible unit, of light. It was already known that light could be characterized, under certain circumstances, as discrete particles called **photons**; according to the Bohr model, each transition involved only that photon whose energy equalled the difference in the electron energy levels. The new theory

explained many experimental results on light emission from hot gases with elegance and simplicity.

By the mid-1920s, scientists had developed most of the modern picture of the atom. Every atom has a fixed number of protons, which specifies its atomic number, and this number uniquely determines which element the atom represents. The protons reside in the nucleus of the atom, while electrons orbit far away, on the scale of the atom. The swarm of electrons is arranged in *shells* of increasing energy levels. Shells are further subdivided into *orbitals* of slightly differing energies. Only two electrons may occupy each orbital, but the total number of orbitals differs for each shell. The innermost shell, of lowest energy, has only one orbital and can contain only two electrons, whereas outer shells can hold more electrons, always in multiples of two. The chemical properties of an atom are determined by the number and arrangement of its electrons. Atoms that are electrically neutral rarely possess a set of fully occupied shells. Atoms engage in chemical reactions in an attempt to fill shells that have available orbitals. Only the *inert gases* (sometimes called the noble gases) of helium, neon, argon, krypton, xenon, and radon have filled orbitals; consequently, they participate in almost no chemical reactions. The strict regularity in the filling of electron shells accounts for the patterns in the Periodic Table; atoms with similar numbers of unpaired electrons have similar chemical properties.

An atom that literally gains or loses an electron, thereby acquiring a net electrical charge, is called an *ion*. Two ions of opposite charge that approach closely can be electrically attracted and thus can sometimes adhere to form a chemical compound; such a bond is said to be *ionic*. The most familiar example of a compound held together by an ionic bond is sodium chloride, ordinary table salt. Most atoms cling fairly tightly to their electrons, however, and the most common type of chemical bond is the *covalent* bond, in which the atoms share electrons.

Many elements have an unfilled outer shell; under the right circumstances they are likely to lose an electron, becoming a positively charged ion. One means by which this can happen is related to temperature; a sufficiently large heating can provide enough energy to an outer electron to liberate it completely from the nucleus. Since the temperature required for this to occur varies and is characteristic of each element, the ionization state of a distant cloud of cosmic gas can provide clues to the temperature of the gas. Another example of the importance of ionization to astronomy is the early universe; during the very early history of the universe most of the hydrogen was ionized. A positively ionized gas is considerably more opaque to light than is a neutral gas, because the free electrons interact with the light and scatter it. This phenomenon places a fundamental limit on how far out, and hence back in time, that we can see.

Chemistry is understood in terms of electron orbitals

Ionized gas consists of positively charged ions and free electrons

Nuclear physics

Atomic nuclei are composed of protons and neutrons

During the 1920s physicists concentrated on atomic theory, arriving, with the help of the new quantum theory, at the model just described. Not much attention was paid to the nucleus. The only elementary particles known were the electron and the proton; it was thus assumed that electrons were present in the nucleus, as well as in the surrounding shells, although all nuclei still had a net positive charge. This model of the nucleus was probably the best that could have been devised at the time; besides, it was aesthetically pleasing to most scientists to think that the universe consisted of two particles of opposite charge. Unfortunately, this simple picture met the fate of many others: new discoveries that contradicted it. In 1932, a new elementary particle was discovered, the *neutron*. The neutron is slightly more massive than the proton and, as its name indicates, it has no net electric charge. It was quickly realized that the neutron was the missing particle of atomic theory; it was the true nuclear partner of the proton.

Neutrons and protons together make up the nucleus of atoms, and are collectively known as **nucleons**. The electrical charge of the nucleus determines the atom's electron structure; hence the number of protons determines the type of element. Two atoms with the same number of protons but different numbers of neutrons are **isotopes** of the same element. Atoms are denoted by a symbolism of the form ${}_p^{n+p}Z$, where Z stands for the one-or-two-letter symbol for the element, p indicates the number of protons (often omitted, since that is always the same for a given element), and $n + p$ is the total number of neutrons and protons. For example, the isotope of carbon that contains 6 protons (making it carbon) and 6 neutrons, for a total of 12 nucleons, is symbolized as ${}_6^{12}\text{C}$. Isotopes occur in different abundances; for most elements one isotope dominates, while the others are relatively rare.

The discovery of nuclear reactions

In retrospect, of course, there had long been clues that the universe was not so simple as to consist of only two kinds of particle. In 1896 Henri Becquerel discovered that a crystal of a uranium compound resting atop a sealed, unexposed package of photographic film left an image. Becquerel had discovered *radioactivity*, the first known **nuclear reaction**. We have already seen that chemistry occurs only among the electrons in the cloud surrounding the nucleus; the nucleus itself is never affected by any chemical reaction. Nuclear reactions, on the other hand, directly involve the nucleons. The nuclei of radioactive isotopes are unstable and emit radiation of some form, which may transform the nucleus. This radiation is of three types: *alpha particles*, which consist of two protons and two neutrons, *beta particles*, which are electrons, and *gamma rays*, which are essentially light rays of very high energy. When a nucleus emits an alpha particle, it *transmutes* into another element, that which has two fewer protons; it also drops in neutron number by two. For example, ${}_{92}^{234}\text{U}$ (uranium 234) is an alpha-emitter; the result is a nucleus of ${}_{90}^{230}\text{Th}$ (thorium 230). Emission of a beta particle also causes transmutation, because when a beta particle is emitted a neu-

tron is converted into a proton; therefore the atom becomes an isotope of another element, that which has one additional proton and one less neutron. Beta decay causes $^{210}_{82}\text{Pb}$ (lead 210) to transmute to $^{210}_{83}\text{Bi}$ (bismuth 210). Emission of a gamma ray, on the other hand, does not change the elemental identity of the atom. Gamma rays may be emitted either on their own, or in conjunction with alpha or beta particles.

The decay of any particular radioactive nucleus is completely unpredictable; however, if a sample of many nuclei is prepared, after a certain time interval called the *half-life*, half the members of the original sample will have decayed. In another half-life interval, another half will decay, leaving only a quarter as many as were initially present; and so forth. Radioactive decay provides an excellent means of dating samples. For example, uranium decays into lead at a known rate; therefore, comparing the ratio of the amount of lead to the remaining quantity of uranium provides an accurate estimate of the time elapsed since the original sample of uranium accumulated. In the crust of the Earth, radioactive elements are most abundant in igneous rocks, those formed by volcanic eruptions. The oldest volcanic rocks on Earth are found in remote regions such as Antarctica, Greenland, and parts of Canada; they provide a lower limit to the age of the Earth, showing that the planet is at least 3.9 billion years old. Similar principles can be applied to determine the age of the Galaxy, but since there are more potential sources of error and observational difficulties in this case, radioactive dating is less reliable than for the Earth. Nevertheless, estimates of the Galaxy's age from radioactive decay are at least consistent with other evidence.

A nuclear reaction even more extreme than radioactivity was identified shortly after the discovery of the neutron. In 1934 Enrico Fermi was attempting to create heavy elements by bombarding uranium with neutrons. He thought he had succeeded, but his interpretation of his data was criticized by Walter Noddack, who suggested that the uranium had, instead, actually split apart. This possibility was taken up by Otto Hahn, Fritz Strassmann, Otto Frisch, and Lise Meitner, who worked on the problem for five years. Frisch and Meitner developed the theory of atomic fission, while the others searched for experimental evidence. In 1939, Hahn and Strassmann succeeded in demonstrating that $^{235}_{92}\text{U}$ would, upon absorbing a neutron, undergo *fission*, the spontaneous splitting of the nucleus into two much lighter nuclei. Only a few very heavy isotopes participate in fission reactions; the most important natural fissionable isotope is $^{235}_{92}\text{U}$. It can split in a number of ways, with the most common yielding barium, krypton, three neutrons, and a great deal of energy.

Uranium, with 92 protons, is the heaviest naturally occurring element. All of its isotopes are radioactive but their half-lives are mostly quite long, up to several billion years; hence uranium is fairly abundant in the Earth's crust. The heavier elements, called *transuranic elements*, are much more unstable, and occur only under very special conditions that typically must be engineered by humans. Of the transuranic elements, the most important is plutonium. Plutonium is very readily fissionable

The half-life is the time required for half of a sample of a radioactive element to decay

Nuclear fission is the splitting of atomic nuclei

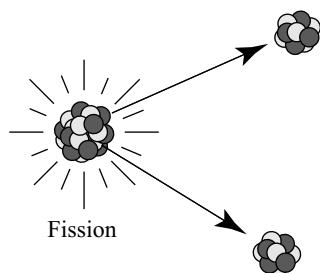


Fig. 4.5 Nuclear fission occurs when a heavy nucleus splits into two lighter nuclei.

Nuclear fusion is the combining of light nuclei into a heavier nucleus

and liberates a great deal of energy when it splits. If it does not fission, it decays by emitting alpha particles. Plutonium is the basis for fission weapons, whereas most nuclear reactors for the production of electrical power use uranium (some use plutonium). Plutonium can be created from uranium 238 by neutron bombardment in a nuclear reactor.

Fission is one kind of nuclear reaction that involves the heaviest atoms. Some of the lightest elements will take part in another kind of reaction, *fusion*, in which two nuclei combine or fuse into a heavier element, with the liberation of various particles as well as *much* energy. Fusion reactions occur only at extremely high temperatures and densities, as the two nuclei must be forced very close together before fusion will occur. Nuclear fusion is an especially important process in the present universe, as it is the source of energy of the stars. The Sun, and other stars like it, shines by nuclear fusion operating deep within its core that converts four hydrogen nuclei (protons) into two neutrons and two protons combined into one nucleus of ${}^4\text{He}$. Humans have been clever enough to learn how to initiate uncontrolled fusion reactions, in bombs, by first raising the temperature and densities to the necessary levels via a fission explosion. We have not yet learned to control fusion reactions as a power source, however, primarily because such high temperatures and densities are extraordinarily difficult to create and maintain on Earth. In the Sun and other stars, fusion occurs in the innermost core, where the tremendous pressure due to the weight of the overlying layers confines the nuclei and creates the high temperatures and densities required.

Discovery of the neutron

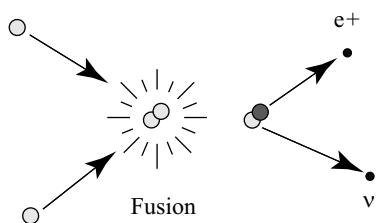


Fig. 4.6 Nuclear fusion occurs when two light nuclei fuse to form a heavier nucleus. Here two protons fuse to become a deuterium nucleus, plus a positron and a neutrino.

Nuclear theory opened up grand new vistas in physics. In the 19th century the only forces known were the electromagnetic force and gravity. It quickly became apparent that neither of these could possibly have anything to do with nuclear reactions. Beta decay was particularly enigmatic, once it was realized that no electrons are present in the nucleus. If that was the case, where did the electron, which was generally ejected from the nucleus with a high energy, originate? Enrico Fermi developed a theory of beta decay in 1934 that introduced the idea that a neutron could be converted into a proton and an electron. In order to satisfy the conservation of momentum, Fermi postulated the existence of a new particle, the **neutrino**. (The neutrino had been suggested previously by Wolfgang Pauli, but Fermi first developed the mathematical theory.) “Neutrino” means “little neutral one”; Fermi gave it such a name because, in addition to being electrically neutral, it had, at most, a very tiny mass. Later theories assumed that the neutrino was massless, but evidence has recently grown that it does have a small mass; as we shall see, this question has significant cosmological ramifications.

The world of modern physics

At the beginning of the 20th century, many of the era’s leading scientists had refused even to accept that atoms existed; less than four decades later, the structure of the nucleus was nearly established. Even the

stars themselves had yielded their deepest secret, the source of their energy. The nucleus itself soon seemed to represent only one aspect of the interactions among particles; after the Second World War particle physics began to emerge as one of the most active subfields of physics. Although the proton, neutron, electron, and possibly the neutrino are the most important of the elementary particles in the present universe, there are many others. As physicists studied *cosmic rays*, which are high-energy rays impinging upon the Earth from space, they found new kinds of particles. With the construction of accelerators, even more particles were discovered. Many of these less familiar particles are unstable; with some very short half-life, they decay into other particle species. As more and more elementary particles were found, physicists realized that there had to be a classification scheme to make sense of them.

Quantum mechanics is the system of physical laws that governs the behavior of the elementary particles, and of nuclei and atoms. It is a formal, mathematical system that developed from the work of many of the greatest scientists of the 20th century, such as Niels Bohr, Max Planck, Albert Einstein, Werner Heisenberg, and Erwin Schrödinger. Quantum mechanics made it possible to sort out the confusing extravagance of particles and to understand their behaviors. The salient feature of any quantum property is that it is quantized; it cannot occur in arbitrary amounts, but only in multiples of a certain inherent value. Electric charge is an example of a quantum property; any particular particle, such as an electron, always has the same, specific quantum of electric charge. Particles possess many quantum properties and may be classified in various ways, depending upon the problem at hand; for now we shall be concerned with only one important property. According to modern particle physics there are two fundamental classes of particle, with the division based upon the *spin* of the particle. The spin of a particle is similar to the spin of a macroscopic object such as a baseball, but with the important difference that it is quantized. Remarkably, the spin of a particle may take either integer or half-integer values; that is, multiples of 0, 1, 2, etc. or $1/2, 3/2, 5/2$, and so forth, of a fundamental quantum unit are permitted, but nothing else. A particle with integer spin is a **boson**, while one with half-integer spin is a **fermion**. The photon has a spin of 1; it is a boson. The electron has a spin of $1/2$; it is a fermion. Spin has a direction as well as a magnitude, and this is also limited to discrete quantum amounts. The number of possible orientations depends on the spin. An electron with spin $1/2$, for example, may be up or down, relative to any particular direction the experimenter might choose. By convention, an up spin is positive, for example, $1/2$ for an electron, while a down spin is negative, $-1/2$ for the electron.

Quantum mechanics is the theory of elementary particles and their interactions

There is a fundamental division of labor between bosons and fermions. The primary duty of bosons is to carry force and energy. Fermions, on the other hand, make up matter. In addition to their different jobs, bosons and fermions have drastically different properties. The most important of these is related to their sociability, in a loose manner of speaking. Bosons are content with one another's company and arbi-

The exclusion principle

trary numbers of them can crowd arbitrarily close together. Fermions, in marked contrast, obey the Pauli **exclusion principle**, a property worked out by Wolfgang Pauli. The exclusion principle is a limitation upon the *quantum state* of fermions, where a state consists of a description of everything that quantum mechanics permits us to know. The state of a particle might include its energy, its spin, whether the spin is up or down, and so forth. According to the Pauli exclusion principle, fermions of the same species that can interact with one another may not simultaneously occupy the same quantum state. The exclusion principle explains why only two electrons are permitted to occupy each orbital around an atom; one has spin up, the other spin down, but otherwise their states are the same. The exclusion principle also demands that fermions cannot crowd together, since interacting fermions must have distinct quantum states. No matter how bizarre it might seem, the exclusion principle controls much of the behavior of matter at the scale of atoms, nuclei, and particles. It is of fundamental importance in the structure of white dwarf stars and of neutron stars, for example.

The first quantum theory was nonrelativistic, but soon quantum mechanics and special relativity were combined into relativistic quantum mechanics. (Quantum theory has yet to be combined with general relativity theory.) This new theory contained a remarkable prediction: the existence of **antimatter**. Every particle has a partner called an *antiparticle*. The antiparticle is, in some respects, the mirror image of the particle, as it has the identical mass; an antiparticle differs from its partner by possessing the opposite sign of electrical charge as well as opposite sign of some other quantum properties. (A neutral particle has a neutral antiparticle.) Only the antiparticle of the electron has its own name; it is called the *positron*. When a particle collides with its antiparticle both are converted to pure energy, in the form of gamma rays. A few particles, most importantly the photon, are their own antiparticles. The universe today appears to be composed entirely of matter, although early on both matter and antimatter were present in great abundance.

Antimatter

No 19th-century scientist would even have dreamed of the menagerie of particles that were known by 1940. Although right now nature might seem to be excessively complicated, there is an underlying simplicity, which is partially understood. The myriad particles interact with one another in various ways, but all of the known interactions can be explained as due to one of only four fundamental forces of nature. These four forces are the **strong interaction**, which holds nucleons together in the nucleus; the **weak interaction**, which mediates nuclear reactions such as fission and beta decay; the **electromagnetic force**, and **gravity**. According to modern theories of particle physics, these four fundamental forces arise due to the exchange of carrier bosons called, for rather obscure historical reasons, *gauge bosons*. The binding between particles by a given force is thus much like the tie between two people playing catch by tossing a ball back and forth between them: the ball carries momentum and energy from one player to the other. The electromagnetic force is particularly well understood; its gauge boson is the

The fundamental forces

particle of light, the familiar photon. Just as the ballplayers can toss a lighter ball farther, the range of a fundamental force is determined by the mass of its gauge boson. The photon is massless; therefore the range of the electromagnetic force is unlimited. Gravity is also carried by a massless boson, the *graviton*, which has so far eluded detection. The weak interaction is mediated by a massive particle; hence its range is limited. The strong interaction has an unusual behavior: as the distance grows, the nuclear force increases. Within the nucleus, the strong interaction has a massive carrier known as the *pion*. At higher energies, in the strange world of *quarks*, the carrier boson is a massless particle known as the *gluon*. Quarks are the fermions from which the nucleons, as well as some other particles, are constructed; gluons hold them together in the nuclei of atoms.

Of all the forces, the strong interaction is, as its name implies, the strongest; its strength exceeds that of the electromagnetic force by about a factor of 100. It would have to be so, or it could not accomplish the job of holding the positively charged nucleus together in the face of electromagnetic repulsion. The weak nuclear force is much weaker than the electromagnetic, by a factor of 10^{11} . But even that huge difference is dwarfed by the full range between the strong interaction and gravity: about 10^{41} . The two nuclear forces operate only over very small distance scales of around 10^{-15} cm, comparable to the size of a nucleus. The electromagnetic and gravitational forces, in contrast, are *long-range*; both diminish as the inverse square of the distance between two particles. We shall not return to the nuclear forces until we study the early universe; here we will describe the other two.

The electromagnetic force is the force that exists between charged particles; it is ultimately responsible for many of the everyday forces we experience. It directly holds ions together in ionic bonds, by the attraction of positive and negative charges. It also causes molecules to stick to one another, because molecules almost always have some distribution of charge even if they are neutral overall. It is the adherence of molecules, through the relatively weak electromagnetic forces between them, which holds together almost all everyday objects, including our bodies. Glues work by causing various molecules to link together. The floor does not collapse under a weight because its molecules are electrostatically bound to one another. Friction is simply the very weak attraction of the surface molecules on one object to the surface molecules on the other object. The electromagnetic force is also responsible for the generation and transmission of electromagnetic radiation, that is, light.

The last of the four fundamental forces is gravity. Although gravity is an incredibly weak force compared to the others, it nevertheless is the most important for the universe as a whole. This is because the nuclear forces are short-ranged, while the electromagnetic force is almost always *shielded*, or reduced, because most things in nature are, overall, electrically neutral. If this were not so there would be enormous forces, as unshielded negative and positive charges attract one another very strongly, much more strongly than they attract one another gravitation-

Gravity controls the evolution of the universe

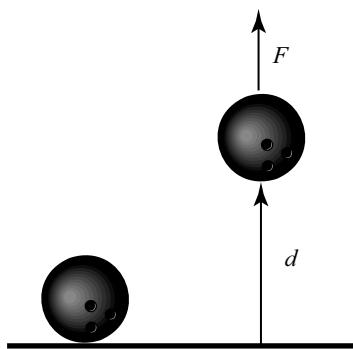


Fig. 4.7 Lifting a bowling ball requires the expenditure of a quantity of work equal to the force applied times the distance the ball is lifted, $W = Fd$.

Types of energy

ally. Huge currents would result, as charges were pulled to one another, until approximate neutrality would quickly prevail. No comparable effects occur for gravity because mass, which plays the role of gravitational charge, is of only one type; there is no possibility for shielding or neutralization by a charge of opposite type. Therefore, over scales larger than approximately 10^{-6} cm, the typical distance between molecules, gravity dominates. It is gravity that shapes the universe we observe, and through most of this book we will be concerned with the gravitational force.

A convenient mathematical way of representing a force is by means of a *field*. A field is a function that fills space and describes the strength of the force at any point. Thus we may speak of the *gravitational field*, a representation of the force of gravity at all points. Similarly, we speak of the *electromagnetic field*, which can itself be broken into an electric field and a magnetic field. More generally, the term *field* can describe any physical entity that has an extension in space, such as the distribution of temperatures in a solid.

Allied with the concept of force are *work* and **energy**. In physics, the quantity *work* is defined very precisely, as the exertion of a force to produce a displacement. Though both force and displacement are vectors (displacement is distance plus direction), work has only a magnitude, not a direction. A quantity that is fully described by its magnitude is called a *scalar*; thus force is a vector, but work is a scalar. Since both the quantities that enter into computing the work are vectors but the result is a scalar, there must exist ways of combining vectors to obtain scalars. In fact, a number of methods can be defined to obtain a scalar output from vector input; work is computed by one very useful method, but it is beyond our scope in this text. In the special case that the force and the displacement are parallel, then the work is just the magnitude of the force multiplied by the distance.

A quantity related to work is **energy**. Energy can be defined as the *capacity to do work*. Examples of energy include *chemical energy*, the energy required to create or break a chemical bond, and *energy of deformation*, the energy required to change an object, such as the fender of a car, into a deformed state. If work is done against gravity to lift a ball to a certain height, the ball acquires **potential energy**. If it fell, it could strike the fin of a turbine and turn it, causing work to be done. The motion of the ball was the direct cause of the turning of the turbine and therefore energy must be associated with motion; energy of motion is called **kinetic energy**. As the ball falls, it loses potential energy and gains kinetic energy. Its kinetic energy is (mostly) converted into work when it hits the turbine. The recognition of the intimate connection between work and energy, by James Joule, was a great step forward in the understanding of **thermodynamics**, the science of energy in general, and heat in particular. **Heat** is a very important form of energy. It is related to the aggregate energy of the random motions of the individual molecules that make up an object, in contrast to what we specifically call kinetic energy, which is a consequence of the bulk

motion of an object as a whole. Heat is used every day to produce work; for example, heat expands the gas in the cylinders of an automobile engine. Although heat and temperature are related, they are not the same thing. **Temperature** is a function of the mean random kinetic energy of molecules, whereas heat depends on such quantities as the individual kinetic energies of the molecules as well as the density of the substance. Thus it is possible for an object to have a very high temperature but relatively little heat energy. The corona of the Sun is a tenuous halo of ionized gas that surrounds the Sun and is visible during solar eclipses; it has a temperature of millions of degrees, but is so thin that its heat content is not extreme. As a general rule, however, higher temperature is associated with a greater quantity of heat energy.

One of the most important laws of physics is the law of **conservation of energy**, which states that *energy is neither created nor destroyed, but is only converted from one form into another*. In classical physics there is a companion to this law, the law of **conservation of matter**, which similarly states that *matter is neither created nor destroyed*. When we study special relativity we shall learn that mass and energy are equivalent, through Einstein's famous equation $E = mc^2$; mass itself is just another form of energy. Special relativity shows that the separate laws of matter and energy conservation must be superseded by a new principle, that of the conservation of matter-plus-energy. Both the law of conservation of matter and the law of conservation of energy can be considered to be individually valid to a high degree under ordinary conditions. However, in cosmology we shall often encounter circumstances that are far from ordinary, so we must keep this grander principle in mind.

Conservation laws

The conservation of energy is also known as the first law of thermodynamics. Since there is a first, there must also be a second. The second law of thermodynamics is one of the most significant laws of physics, and one of the least understood. Many equivalent statements of the second law exist; for now let us give the version presented by Rudolf Clausius in the middle of the 19th century: *No cyclic process exists whose sole effect is to transfer heat from a cooler to a warmer body*. There are many devices, such as refrigerators and heat pumps, which transfer heat from cooler to warmer bodies, but this process is always accompanied by the exhaust of waste heat. The second law denies the possibility of perpetual-motion machines. Some energy is always dissipated into waste heat in any real, macroscopic process and for this reason no machine, no matter how clever or carefully designed, can ever run forever without an input of energy. No perpetual-motion machine has ever been built. Every one that has been claimed has been found wanting upon close examination. Some have been outright frauds; others were so carefully balanced that they could operate for a very long time, but not indefinitely. A notorious device of recent years whose inventor claimed it to produce *more* energy than it consumed was shown to be nothing but a simple power converter, and an inefficient one at that. (A machine of this kind would also violate the first law of thermodynamics.) The

The concept of entropy

second law has always triumphed, no matter how ingeniously humans have tried to circumvent it.

More modern versions of the second law connect this inevitable dissipation of energy to an increase in **entropy**. Precise, mathematical definitions of entropy exist but, loosely speaking, entropy is a measure of the *disorder* of a system; the higher the entropy, the greater the disorder. The second law can be restated as *in any process, the overall entropy increases, or at best remains the same*. Since an ordered system has a greater potential to do work, an increase in entropy is accompanied by a reduction in available energy. The second law does not deny the possibility of the *existence* of order, however. Order can always be created locally by the consumption of energy. As an example, biological systems—living creatures—represent highly ordered states, perhaps the most highly ordered in our region of the universe. Nevertheless, their mere presence is not a refutation of the second law; on the contrary, modern research on the theory of ordered systems indicates that dissipation is *required* for complex, ordered states to arise naturally. But there is a price for order, and that price is the conversion of available energy into waste heat whose capacity for useful work is greatly diminished. Biological entities obtain their energy ultimately from the Sun, or in a few species from geothermal energy. Like any other macroscopic process, life results in an overall increase in the entropy of the universe. Neither are artificial processes immune to the second law; energy must be expended to support manufacturing and transportation, with the inevitable consequence that entropy increases and the Earth's supply of utilizable energy is reduced.

It may seem that the second law exists only to frustrate human attempts to get something for nothing. A consistent system of thermodynamics could be developed without it; yet it is always confirmed, not only in experiments but in the realities of engineering and everyday life. The second law seems intimately related, in ways which we cannot yet fully comprehend, to the earliest moments of the universe, as well as to its ultimate fate. The second law appears to determine the *arrow of time*, the relentless march of time in one direction only. The second law of thermodynamics may be one of the deepest, most fundamental rules of the universe.

The third law of thermodynamics essentially completes the foundation of the system. (A fourth law, called the zeroth law, provides a statement that thermal equilibrium is possible.) The third law is a consequence of the observation that cooling to very low temperatures is difficult, and becomes more difficult as the temperature is lowered. The lowest possible temperature is *absolute zero*. The third law of thermodynamics states that *absolute zero can never be attained, but only approached arbitrarily closely*.

Waves

The material world of particles and atoms seems concrete and familiar; yet just as important to cosmology is the incorporeal world of waves. A *wave* is a disturbance in some quantity that propagates in a regular way. We are all familiar with water waves, ripples in a body of water that move across the surface, leaving it undisturbed after they pass. Waves carry energy with them as they travel, as anyone who has ever stood in the surf should realize. Waves are characterized by maxima called *crests* and minima called *troughs*. The maximum displacement from the undisturbed position is the *amplitude* of the wave. The number of crests that pass an observer in a specified unit of time is the *frequency* of the wave, while the distance from one crest to the next is the *wavelength* of the wave. For a pure or *monochromatic* wave, the wavelength is a well-defined constant. In general, however, an arbitrary wave is a *superposition* of many pure waves, and the wavelength is not so easy to define. The distribution of frequencies in a superposed wave is called its **spectrum**. The energy carried by a wave is related to its frequency; generally, the higher the frequency, the greater the energy transmitted.

Important examples of waves include sound waves, which are oscillations in the pressure of a gas or liquid, and water waves, which are displacements of parcels of water. (Sound waves can travel in water, but they differ from what is defined as a water wave.) Sound and water waves are examples of waves that require a *medium* for their propagation. The particles of the medium move very little, whereas the wave can move, and transmit energy, over great distances. Water waves easily travel across the Pacific Ocean, while similar waves in the Earth's atmosphere can circle the globe. For astronomy the most important type of wave is the electromagnetic wave, that is, light. Light differs from the other waves described here in that it does not require a medium for its transmission, but otherwise its properties are similar to those of any wave.

Waves have several unique behaviors. They can undergo *reflection*, partially or completely, when they strike a surface; that is, part or all of the wave train turns back and travels in the opposite direction. When a wave passes from one type of medium to another, *refraction*, a change in the wave's speed and thus its direction of motion, occurs. An example of refraction in water waves is the bending of the waves as they move from deeper water into a shallow inlet. When a wave passes through an opening that is comparable in size to its wavelength, it undergoes *diffraction*, the bending around the obstacle. Diffraction of sound waves enables voices to be heard through an open door even when the speaker is not aligned with the doorway.

When two waves of similar type pass through one another, **interference** can occur. Two crests or two troughs may meet and reinforce one other, creating *constructive* interference and resulting in a greater displacement at that point than is present in either individual wave. If a crest and a trough meet, the result is *destructive* interference, in which

Properties of waves

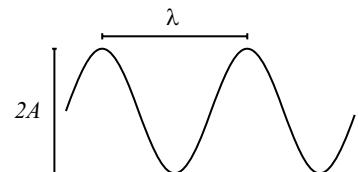


Fig. 4.8 Schematic illustration of a monochromatic wave. A is the amplitude, while λ is the wavelength. A pure monochromatic wave would extend in both directions indefinitely.

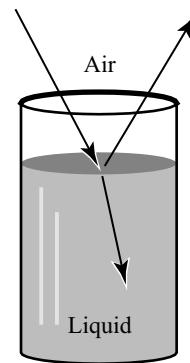
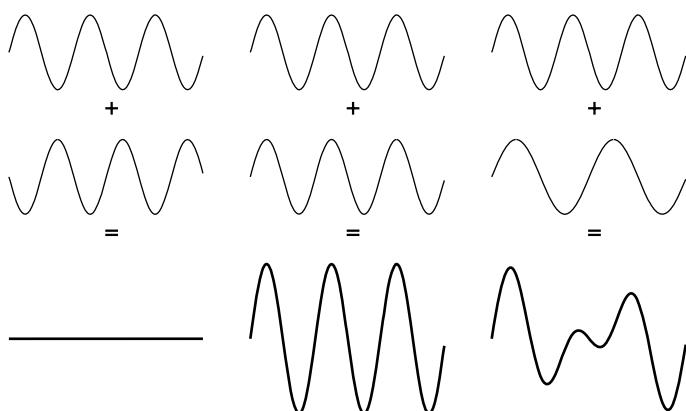


Fig. 4.9 Wave reflection and refraction occur at the boundary between two media through which the wave propagates with different speeds. The angle of refraction depends upon the change of the wave's speed as it enters the new medium; the angle of reflection is equal to the angle of incidence.

Fig. 4.10 Wave interference. Two waves that pass through one another reinforce or cancel, partially or wholly. The net wave (**bold line**) is the sum of the amplitudes of the interfering waves at any point in space. In the first column, two waves of equal wavelength and amplitude, but precisely out of phase, cancel exactly. In the second column, two waves of equal wavelength and amplitude and of the same phase add coherently. The third column depicts a general case with two waves of different wavelengths. Interference is difficult to visualize, but it can sometimes be directly observed in water waves.



the net displacement is reduced, sometimes even exactly canceled. If the waves are linear, the interference may be computed at each point simply by summing the two amplitudes (positive for crests, negative for troughs) at that point. For nonlinear waves, this simple law does not hold, but interference still occurs. One important interference effect occurs when two waves superpose to form a pattern of alternating light and dark bands, called *interference fringes*. An example is provided by monochromatic light passing through two closely spaced slits and projecting onto a screen. The wave crests from the two slits alternately reinforce and cancel one other, creating a characteristic pattern. The appearance of these interference fringes is a definite indicator of the wave nature of light.

One of the most important consequences of wave properties, from the point of view of astronomy, is the **Doppler effect**. The Doppler effect is familiar to everyone when it affects sound waves. As a train approaches a grade crossing, the driver waiting in his car hears the pitch of the whistle rise. After the train passes, the pitch drops. If the train and the driver were at rest with respect to one another, the sound waves from the whistle of the train would move outward in a roughly spherical pattern; the constant wavelength between successive crests would determine the fundamental pitch of the whistle. When the train is approaching, however, it is moving in the same direction as are the wave crests which reach the driver; thus each successive emitted crest follows the previous one at a shorter interval than if the train and driver were mutually at rest. Conversely, as the train recedes, it is moving opposite to the direction of motion of the wave crests reaching the driver's ears, so successive waves arriving at the driver's position are spaced at longer intervals than they would be if the train and driver were at rest.

We can illustrate this phenomenon more concretely. Suppose you decided to learn to play tennis, but you could not find a human practice partner patient enough to put up with your attempts to bat the ball

The Doppler effect

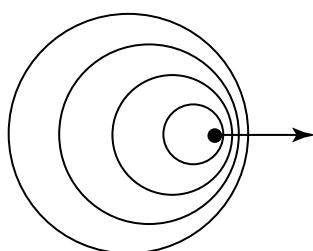


Fig. 4.11 The Doppler effect. When the source is approaching the observer, the wave crests bunch up, resulting in a shorter observed wavelength. When the source is receding from the observer successive wave crests are stretched, giving a longer observed wavelength.

around. You might then use a device similar to a miniature cannon, which shoots tennis balls at a constant rate, as seen by you when you stand at rest near the rear of the court and watch the balls fly past. If you ran toward the cannon, the interval between the balls you would encounter would be shortened, because each successive ball would now have less distance to cover before meeting you. Conversely, if you ran away from the cannon, each ball would have to make up the extra distance caused by your recession before it could reach you, and thus the interval between balls would, as seen by you, increase. The individual tennis balls could correspond to the crests of a wave, the cannon to any kind of source. In this situation, the receiver, you, is moving, but there is still a Doppler effect, so clearly it cannot depend upon whether the source or the observer is moving. This example also demonstrates that the Doppler effect can occur for *any* kind of periodic phenomenon.

The Doppler effect is a consequence of the *relative motion* between the source and the observer. The effect depends only upon the nature of waves and upon the motion of the source relative to the receiver, and thus this phenomenon affects light waves in exactly the same way as sound waves. If the source is approaching the observer, relative to the observer's frame of reference, the light waves bunch up and are shifted toward higher frequencies; this is a **blueshift**. If the source is receding from the observer the light waves shift to lower frequencies, resulting in a **redshift**. The formula for the Doppler shift of light is, for relative speeds v much less than the speed of light c :

$$z = \frac{\lambda_{\text{rec}} - \lambda_{\text{em}}}{\lambda_{\text{em}}} = \frac{v}{c}, \quad (4.1)$$

Redshift is an extremely important astronomical phenomenon

where z is the shift, λ_{em} is the wavelength in the frame of the emitter, and λ_{rec} is the wavelength in the frame of the receiver. A negative value of z indicates a blueshift; positive z gives a redshift.

On Earth, the Doppler effect has found numerous applications, including the radar speed guns by which a highway patrolman may measure the speed of approach of an automobile. (In a radar system, a transmitter emits radio waves that reflect from a target and return to their source.) Astronomy depends particularly heavily upon the Doppler effect. For stars, nebulae, and nearby galaxies, the Doppler shift can tell us how fast the object is moving toward or away from the Earth. The major shortcoming of this technique is that it cannot give us the absolute velocity, but only its radial component. If an object is moving transversely to the Earth, then it is neither approaching toward nor receding from us, and there is no Doppler shift.² Consequently, we cannot detect the transverse component of the velocity by means of a Doppler shift. Even with this limitation on our knowledge, however, considerable useful information can be determined. For example, if we measure both

²There exists a transverse Doppler shift that is a consequence of time dilation at relativistic speeds, a topic which will be covered in Chapter 7. But almost all objects move at small speeds compared to c , and the relativistic transverse Doppler shift is insignificant and usually unobservable.

a redshift and a blueshift from different regions of an object, we can conclude that the object is rotating, and we can measure its rotational speed.

The nature of light

Light: particle or wave?

The most important wave to astronomy is *light*. We spend our lives immersed in light, but how many have wondered what it is? The nature of light was argued and debated for centuries, but only during the past 300 years, since the development of experimental science, has any significant progress occurred. Newton performed some of the most important early experiments; it was Newton who showed that white light was a combination of all colors. Newton also studied a specific kind of interference fringes, now called *Newton's rings*, which occur when a glass plate with a very slight curvature is placed over a flat plate, and the whole assemblage is illuminated from beneath. Oddly, although interference fringes are an unmistakable signature of a wave, the rings were not recognized as such. Newton believed quite firmly that light was corpuscular, consisting of a stream of particles, since he could not accept that a wave could account for the apparently straight and narrow propagation of a beam of light. Other scientists of the time, most prominently Christian Huygens, were equally convinced that light was some kind of wave, but this faction held almost equal disregard for what experimental evidence then existed. The matter seemed finally resolved in 1803, when Thomas Young passed light through two very narrow slits in a solid plate and obtained interference fringes. A century later, however, Einstein revived the corpuscular theory of light, but in a form which Newton would not have recognized, and which he probably would have disliked. Light can show both particle and wave natures, though only one at a time. We shall usually need only of its manifestation as a wave; but occasionally the particle nature of light will be important, especially when we study the early universe. For now let us concentrate on the wave properties of light, with some allusions to its particle manifestation.

The electromagnetic spectrum

Visible light is a specific type of *electromagnetic wave*. Electromagnetic waves are traveling disturbances in the electromagnetic field. Unlike other kinds of wave, they do not require a medium for their propagation, although this important fact was not understood until early in the 20th century. All electromagnetic waves are of the same nature, differing only by their wavelength. The full range of such waves is called the *electromagnetic spectrum*. For the convenience of humans, the electromagnetic spectrum is divided into *bands*, or groups of frequencies. At low frequencies, we call the waves *radio waves*. Progressing to higher frequencies, we have *microwaves*, *infrared radiation*, *visible light*, *ultra-violet radiation*, *X-rays*, and finally, at the shortest frequencies, *gamma radiation*. The difference in names is due to the separate discoveries of different portions of the electromagnetic spectrum before it was recognized that all these waves were of the same kind. The division into

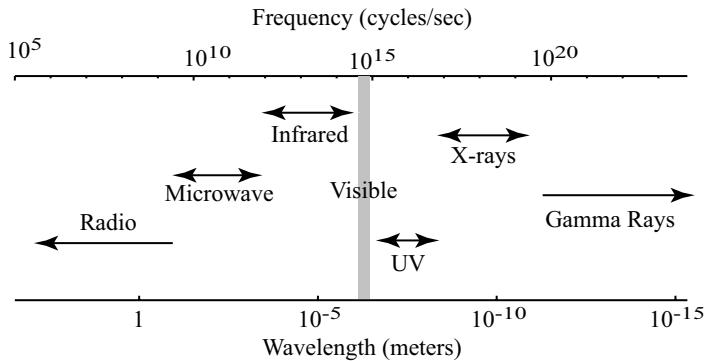


Fig. 4.12 The electromagnetic spectrum. The scales show the wavelength in meters, and the wave frequency in cycles per second. The high-energy, short-wavelength region is the regime of gamma rays and X-rays. The low-energy, long-wavelength portion is the radio region. Visible light, that part of the spectrum to which our eyes respond, is located between the ultraviolet and the infrared. The division into bands is conventional and somewhat arbitrary.

bands is also quite arbitrary, and has no particular physical significance. The most obvious subdivision is visible light, which is defined as that band that the human eye can detect. However, even here there is some ambiguity, as different people can see slightly different ranges; in fact, people who have had the lenses of their eyes removed can see into the ultraviolet.³ It is common to employ the word “light” as a generic term for all of the electromagnetic waves, and we shall do so unless there is some need to distinguish one band from another. However, visible light is not qualitatively different from any other part of the electromagnetic spectrum.

The relationship between wavelength and frequency for an electromagnetic wave traveling in a vacuum is very simple:

$$\lambda\nu = c, \quad (4.2)$$

where as usual λ is the wavelength, and ν is the conventional symbol for frequency. In this formula, c is a constant of proportionality between the two quantities. It has units of speed and turns out to be the speed of motion of the wave in the vacuum; it is called the *speed of light*. All electromagnetic waves travel at this same speed in a vacuum. In a medium, however, a group of electromagnetic waves initially traveling together will traverse the medium at different speeds, always less than c ; this phenomenon is called *dispersion*. When white light, which consists of a superposition of all the wavelengths in the visible band, is passed through a prism, the different wavelengths travel through the glass with slightly different speeds. This causes them to refract differently at each of the two surfaces they cross. As a result, the prism breaks white light into its monochromatic components. In the field of *spectroscopy*, the analysis of spectra, the superposition of all wavelengths is called the *continuum*.

The speed of light in a vacuum is enormous in comparison to almost any other speed we can imagine; but it is finite, and that has important

³The lens absorbs ultraviolet rays and prevents them from striking the retina, which might be damaged by the higher-energy light. Exposure to ultraviolet light has been implicated in the development of cataracts.

implications for cosmology. When we look at a star, we see that star not as it is now, but as it was when the light departed from it. Looking into space is equivalent to looking back in time. The distance light travels in a year is called one **lightyear**. In MKS units, the speed of light c is $2.998 \times 10^8 \text{ m s}^{-1}$; hence a lightyear is 9.5×10^{12} kilometers, or about 6×10^{12} miles. Notice that the lightyear is a unit of distance, not of time.

Radiation is the general term for the emission of energy from an object, often in the form of a wave. The word is frequently applied to the emitted wave or particle itself, as in the expression “ultraviolet radiation.” Nuclear radiation, which we have already discussed, may consist of particles, specifically helium nuclei (alpha) or electrons (beta); only gamma rays are actually photons. Charged particles radiate electromagnetic waves when some of the particles’ energy is converted into photons. One example is a transmitting antenna, which converts current (moving charge) into electromagnetic radiation. In this case, some of the kinetic energy of the charged particles is transformed into radiation.

Thermal radiation is produced by all dense substances

One of the most important sources of electromagnetic radiation in nature is *thermal radiation*. In any substance with a temperature greater than absolute zero, the constituent particles (atoms or molecules) vibrate, jiggle, and possibly rotate. Energy levels are associated with those overall motions; macroscopically, the collective energy of these random motions is what we experience as heat. A portion of this heat energy is converted into photons and radiated away. The spectrum of thermal radiation from an arbitrary object can be quite complex, depending upon such variables as the composition of the object, its shape, how much external energy it is capable of absorbing, and so forth. The only general rule is that the hotter the object, the higher the energy of the photons it emits. We all know this from everyday life. A heated iron emits no visible light, but glows brightly when photographed with film sensitive to the infrared. The coils of an electric stove set on high glow with red light; red is the lowest-energy visible light. The stove also emits a great deal of infrared radiation, which cooks the food, but it is hot enough that some of its emission is in the visible. Hotter objects, such as a very hot poker, emit more and more in the visible until they emit all visible wavelengths, and thus appear white. Still hotter objects acquire a bluish color, as their emission shifts into the higher-energy visible and beyond to the ultraviolet.

Blackbody radiation represents ideal emission from a substance in thermal equilibrium

There is one extremely important special case in which thermal radiation is easily predictable. This is the thermal emission from a *perfect absorber*, called a *blackbody*. By definition, a perfect absorber is also a perfect emitter. Radiation from such an object is called **blackbody radiation**. Of course, a perfect blackbody is an idealization, but close approximations abound, even on Earth. One excellent approximation is *cavity radiation*. An example of such a heated cavity is a pottery kiln. As the kiln heats, its interior fills with thermal radiation emitted by the walls. Since the temperature of the walls surrounding the cavity is the same, the emission and absorption of energy within the cavity must

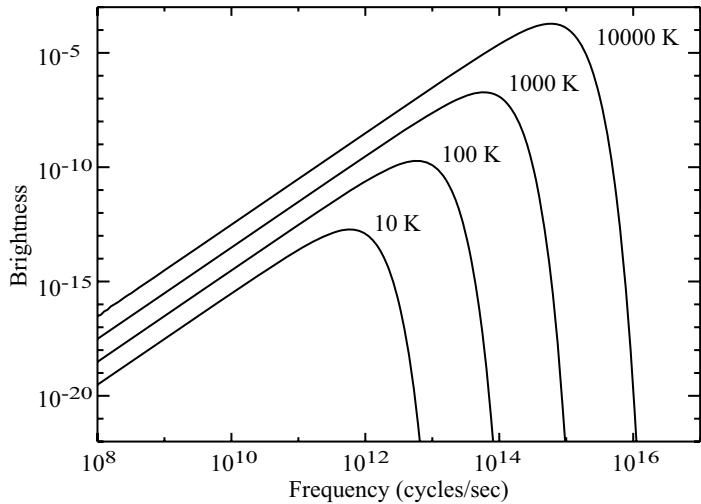


Fig. 4.13 Representative spectra of blackbody radiation for different temperatures. The shape of each curve is the same; only the magnitudes and positions differ. The peak frequency depends only upon the temperature of the blackbody emitting the radiation. Blackbodies emit significant visible light only when their temperatures are greater than about 1000 K. (Compare to Figure 4.12 to determine where the peak wavelength lies in the electromagnetic spectrum.)

come into balance, regardless of the nature of the kiln walls. This is the key characteristic of blackbody radiation: it represents a state of **equilibrium**, or balance, in the photons. If we drill a small hole in one wall of the cavity and sample some of the radiation within, we will find that the shape of its spectrum does not depend upon the configuration or composition of the walls, but *only* upon the temperature. The spectrum rises to a maximum intensity at a certain wavelength, then falls back down toward zero emission. Moreover, the wavelength (or, equivalently, the corresponding frequency) at which the peak of the spectrum occurs is uniquely correlated with the temperature; from only this single datum, we can determine the temperature of the radiation. Specifically, the peak wavelength of the blackbody spectrum is inversely proportional to the temperature of the emitter. The formula relating the peak of the spectrum to the temperature is called the *Wien displacement law*, and is given approximately by

$$\lambda_{\text{pk}} \approx \frac{0.29}{T} \text{ cm}, \quad (4.3)$$

where the temperature T is on the kelvin temperature scale which sets its zero point at absolute zero. The surface of the Sun is a close approximation to a blackbody. Its surface temperature is about 5,800 K, and its spectrum peaks in the visible range at a wavelength of approximately 5×10^{-5} cm.

A very important property of blackbody radiation is that the shape of the spectrum is always the same; it is simply shifted to shorter wavelengths (higher frequencies) at higher temperature. As the temperature of a blackbody is increased, the quantity of energy radiated per unit surface area goes up as the fourth power of the temperature, that is,

$$\mathcal{E} \propto T^4. \quad (4.4)$$

The peak wavelength of a blackbody spectrum is inversely related to temperature

Blackbody radiation fills the universe

Stars are approximately blackbodies, so a hot star with a surface temperature twice that of the Sun would be radiating 16 times as much energy per unit surface area. If that hot star had the same surface area as the Sun, its total emitted energy would correspondingly be 16 times as great. Because such hot stars are typically considerably larger than the Sun, their total energy output is usually much greater than the Sun's.

Blackbody radiation is especially important to cosmology. As we continue our study, we shall discover that the universe is filled with such radiation, with an associated temperature of a mere 2.7 degrees above absolute zero. The presence of this background radiation indicates that the universe was once much hotter, providing direct evidence for what is known as the big bang. How this background radiation originated, and what it can tell us about the early universe, is a story to be developed in later chapters.

At the end of the 19th century, the explanation for the spectrum of blackbody radiation had stymied every great physicist who had worked on it. The peaking and overturning at high energies could not be predicted by the classical laws of thermodynamics. The lower-energy portions ascending up to the peak could be explained by the physics of the era, but the resulting classical formula continued to rise indefinitely, predicting infinite energy at the shortest wavelengths! Since this was obviously impossible, it was called the *ultraviolet catastrophe*.

In 1900, Max Planck presented a formula that fit the data nearly perfectly. Planck had set out to find a theoretical explanation for blackbody radiation. He tried many possibilities, but the breakthrough came when he made the assumption that radiation could be emitted and absorbed only in discrete units. The explanation of blackbody radiation was the first hint of quantum mechanics, a theory that did not develop fully for another twenty-five years. The earlier classical formula worked reasonably well for low energies, where the quantum nature of the light was not very important, but failed at high energies, where only quantum effects could explain the data.

Albert Einstein made another step forward in 1905 when, in an effort to explain a puzzling experimental result on the emission of radiation from metals, he postulated that light acts not only like a wave, but sometimes like a particle; the photon is the quantum of light. This completed the understanding of the blackbody spectrum, since it became clear that Planck's quantum of energy was the photon itself. Although the photon is massless, it (and other massless particles) still transports energy. Each photon carries an amount of energy given by the formula

$$E_\nu = h\nu, \quad (4.5)$$

Light is emitted as discrete bundles called photons

where h is a constant called *Planck's constant*, and ν is the frequency of the corresponding wave. A single photon carries one quantum of energy; hence it is associated only with a single wavelength and frequency, that is, a monochromatic wave. According to the laws of quantum mechanics, light will reveal either its wave or its particle nature in a given experiment, but never both at once.

Advances in the new quantum theory quickly led to another triumph in the understanding of light. When light from a radiating sample of tenuous gas is analyzed, it will be found to consist of bright, narrow lines; such a spectrum is called an *emission spectrum*. Because it consists of distinct spectral lines, this type of radiation is often called *line radiation*. Bohr's work on the quantization of electron orbitals in atoms provided the explanation for this form of electromagnetic radiation; it originates from the *quantum transitions* of the orbiting electrons. As we have discussed, each electron bound to a nucleus must have a well-defined energy, specified by the orbital it occupies. Under certain circumstances, the electron may drop into an orbital of lower energy, emitting a photon in the process. If we define E_i as the energy of the initial orbital, and E_f as the energy of the final orbital after the transition, then the frequency of this photon is obtained from equation (4.5) above:

$$|E_f - E_i| = h\nu. \quad (4.6)$$

Electrons in atomic orbitals emit or absorb line radiation

The orbitals cannot have arbitrary energies, and therefore this difference is always some discrete amount, which depends upon the transition. Obviously, the lower-energy state must be available or the transition cannot occur; recall that only two electrons may occupy each orbital, though several orbitals may make up a shell, and electrons can certainly jump from one shell to another. Since the exact electron configuration is specific to each particular element, the transitions permitted to the electrons depend upon which type of nucleus they orbit. Moreover, because of various rules of quantum mechanics, it may turn out that for a given atom, some transitions are much more probable than others while some are almost forbidden, even if those orbitals are available. Each atom thus emits a unique spectrum of frequencies that is so characteristic that every element may be identified from its spectrum alone.

The inverse process also occurs; an electron may absorb a photon and be boosted from a lower-energy orbital into a higher-energy orbital. Because this can occur only if a photon of exactly the right energy happens to be available, it is also highly specific and characteristic. The element absorbs exactly those frequencies which it would emit in the opposite process. If white light strikes a collection of atoms in the gaseous state, each atom will absorb photons of precise frequencies, and when the light that has passed through the gas is analyzed with a prism, the missing frequencies will be unique to the specific element. Such a spectrum is called an *absorption spectrum*.

The spectrum of the Sun shows a forest of absorption features called Fraunhofer lines, after their discoverer Joseph von Fraunhofer. Atoms in the relatively cool outer layers of the Sun absorb some of the photons generated from deeper, hotter layers. However, this was unknown when the lines were first resolved. (The corresponding emission lines are much more difficult to detect and were not found until two decades after Fraunhofer's death.) The realization came to Gustav Kirchoff and Robert Bunsen, over the period between 1855 and 1863, that these lines could be identified with laboratory spectra of Earthly gases. This stun-

A line spectrum provides a unique fingerprint for an element

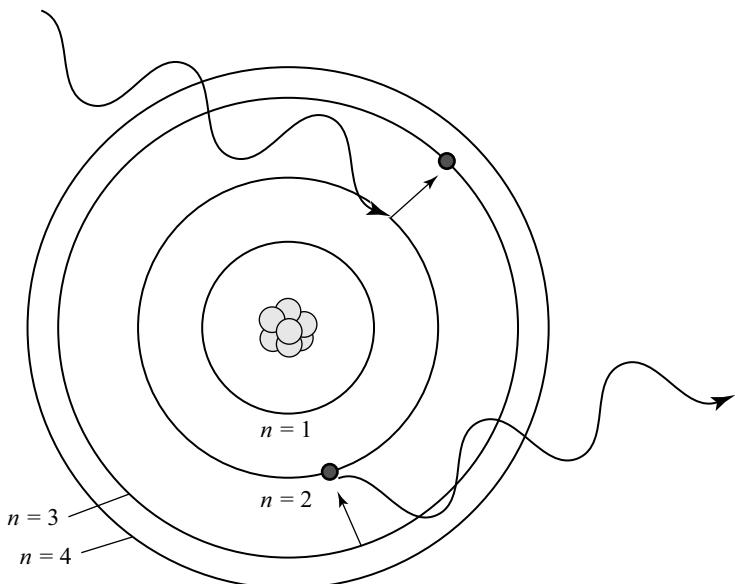


Fig. 4.14 In the Bohr model of the atom, electrons surround the central nucleus in specific orbitals that correspond to particular energy levels, labeled by number n . When an electron makes a transition from a higher energy level to a lower level, a photon is emitted with precisely the energy difference between the levels. Similarly, an electron can jump from a lower to a higher energy level if it absorbs a photon with exactly the required energy.

ning discovery made possible astronomical spectroscopy and paved the way for modern cosmology. Until then, most astronomers had believed that it would never be possible to determine the chemical composition of heavenly objects. But by the 1920s, most of the time of all large telescopes was devoted to spectroscopy, as is still the case today. The demonstration that the Sun and the stars were made of the same elements as were found on Earth was a powerful vindication of the Copernican principle.

Emission, absorption, and thermal radiation can coexist in the same spectrum. When an astronomer photographs the spectrum of an object such as a distant galaxy, she will find lines superimposed on a continuum. Most of the continuum is thermal radiation from the object, while the line radiation consists of discrete, resolvable transitions that are specific to the particular elements present in the object. For example, a portion of the atoms in a cloud of interstellar gas might be directly excited by the light from a bright star embedded in the cloud. The spectrum of such a cloud would show emission lines, which by their frequencies and strengths would reveal the kinds and abundances of elements present. The background due to the thermal radiation of the cloud as a whole would provide an estimate of its temperature.

How brightly they shine

Almost all the information we can gather about the universe and its contents comes from the photons we detect. Astronomy is an *observational* science, as opposed to an experimental science; we cannot arrange controlled experiments to study the universe as a whole, but can only

observe it. It is worthwhile, therefore, to review briefly a few of the basic quantities that arise when measuring the light from the sky.

Stars, and other astronomical objects, give off light. The total amount of electromagnetic energy emitted per unit time (the *power*) by a source is called the **luminosity**, generally symbolized as L , and in astronomy often expressed in terms of the *solar luminosity* L_{\odot} . When an object such as a star shines, light travels outward from all points on its surface. The luminosity is never directly measured. At the Earth we intercept only a portion of the total radiation emitted by an object; only that small fraction of this energy that strikes a detector can be measured. Most people are aware that the brightness of a source goes down with distance. This can be made mathematically exact by imagining a spherical surface surrounding the star, at some distance R from it. Energy must be conserved; if we consider photons that are traveling through the nearly empty space around the star, we can ignore absorption or other losses of energy. From energy conservation, the total amount of luminous energy crossing such an imaginary sphere is the same as that which was emitted at the surface of the star. Since the surface area of a sphere increases as the radius squared, the energy per unit time crossing such surfaces at greater and greater distances from the star must decrease as the inverse of the square of the distance from the star. This argument is succinctly expressed mathematically as

$$\text{brightness} = \frac{L}{4\pi R^2}. \quad (4.7)$$

Brightness is also called the energy *flux*, and it specifies the power per unit area. The distinction between luminosity L and energy flux is that the former refers to the *total* power emitted over the entire source, whereas the latter designates the power received per unit area at a detector located anywhere outside the source.

The relationship between distance and brightness (flux) is very important to astronomy, and especially to cosmology. Astronomers can measure the energy flux, but unless we know the distance to the star we cannot compute its luminosity. Conversely, if we know from a study of nearby stars that a certain type of star has a characteristic luminosity, we can compute the distance to any star of that type by measuring the flux of radiation we receive from it. Such a measurement of distance is often called the *luminosity distance*, in order to emphasize how it was obtained and to indicate the errors that might occur.

Of course, the assumption of no loss (or gain) of energy between us and the source is a major restriction. Space is indeed quite empty, but between us and even nearby stars there is a lot of it, so that even a very small density of matter means that the light which reaches us may have been diminished by more than just the geometrical factor in equation (4.7). Any decrease of flux due to interactions with matter is called *extinction*. The amount of extinction varies in complicated ways with such factors as the quantity and type of intervening matter and the frequency of the light. Because so many unknown factors play a role, constructing a model for the dimming of light by extinction is often not

The inverse square law relates apparent brightness to intrinsic luminosity

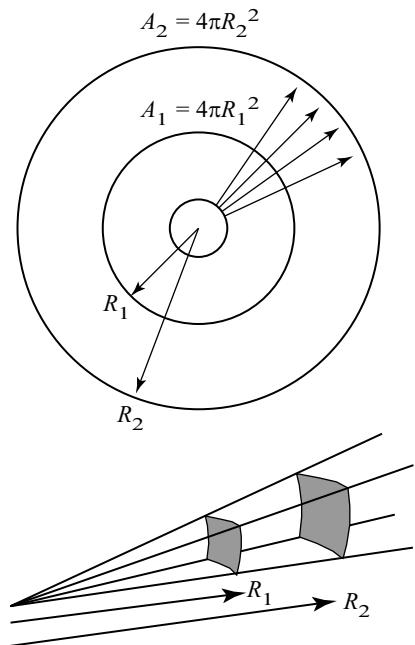


Fig. 4.15 Because the surface area of a sphere increases as the square of its radius, the power per unit area (flux) of light emitted by a point source diminishes as the light travels into space.

The inverse square law provides a way of obtaining distances to astronomical objects

all that easy, but if we are to be able to use luminosity distances with any degree of confidence, it is necessary to account for extinction. This adds considerably to the difficulty of measuring cosmic distances, but the luminosity distance is nearly our only possibility for gauging the farthest reaches of the universe, so we have no choice but to do our best with it.

In addition to the luminosity, the mass of an object is another quantity of great interest in astronomy. We cannot construct gigantic balances to measure the mass of a star or galaxy directly, so how can it be determined? We know that the mass of an object determines the gravitational force it exerts on other objects. For the Sun, we can easily measure its gravitational influence to very high precision, and from this we learn that the Sun has a mass of about 2×10^{30} kg. The mass of the Sun, in whatever stated units, is called one *solar mass*, denoted by M_{\odot} . This mass is so large that it may be difficult to comprehend; for comparison, the mass of a typical human is roughly $3 \times 10^{-29} M_{\odot}$. This is close to the ratio of the mass of a human to that of a single proton.

Measuring stellar masses

We cannot readily determine the mass of an isolated, distant star, but many, perhaps as many as half, of all known stars are members of binary systems, a system of two stars that mutually orbit one another. From Kepler's third law, we can determine the sum of the mass of the two objects; if we can also measure the force between them, we can calculate the masses of the individual objects. By such means astronomers have found that stars range in mass from approximately twenty percent of the mass of the Sun to over fifty solar masses. In addition to binary systems, stars also occur in large groups. It is difficult to apply Kepler's and Newton's laws to such clusters, because the equations of Newtonian gravitation can be exactly solved only for two bodies. All we can obtain from star clusters are statistical properties but, again using Newton's laws, those statistics can provide a good estimate of the cluster's aggregate mass. If we can also observe how many, and what types, of stars are present, we can estimate the mass in luminous objects (usually mostly stars) of the cluster from our knowledge of the properties of other stars of the same types. We shall return to this topic in detail in later chapters.

Where are we?

A scale model of the solar system

The solar system consists of the Sun, a smallish star resident in the suburbs of an average galaxy, and all the lesser objects that are gravitationally bound to it. The Sun dominates its system completely; the second-largest object, Jupiter, has only 0.096% the mass and 2% the diameter of the Sun. There are nine planets and innumerable smaller bodies. We can construct a scale model of the restricted solar system, consisting only of the Sun and the nine planets, to make it easier to grasp the scale of the system. We are going to need a great deal of room to accomplish this, since the solar system is very large and very empty.



Fig. 4.16 A spiral galaxy, NGC 3370, as observed with the *Hubble Space Telescope*. (A. Riess, Hubble Heritage team, STScI/NASA.)

As a start, suppose that the Sun were the size of an orange. The Earth would then be about the diameter of a small BB pellet (1 mm) at a distance of 11 meters from the orange. The Moon is 0.25 mm in diameter and is located about an inch (2.5 cm) from the Earth. Jupiter is about 1 cm in diameter and resides 60 meters, over half the length of a football field, from the orange. Tiny Pluto is only 0.2 mm in diameter, and its mean distance from the orange is 430 meters, about four football fields. Yet even these staggering distances are just down the street compared to separations in interstellar space. The nearest star to the Sun, at a distance of 4.3 lightyears, is Alpha Centauri, a star (more precisely, a stellar system) visible only in the Southern Hemisphere. On our scale model, Alpha Centauri is about 3000 km from the orange. Interstellar distances are really too large to be comprehended by human intuition, yet they are still small compared to the scale even of the Milky Way Galaxy. It is only through the symbolism of mathematics that we are able to understand the nature of the cosmos.

As far as we can tell, essentially all stars occur within **galaxies**. Galaxies are large clusters of stars, gas, and dust that make up the fundamental population of the universe. Galaxies are divided into three major categories. *Spiral galaxies* are great disks of stars, with grand patterns of spiral arms threaded through them like the fins of a pinwheel. The spirals themselves cannot be rigid objects or they would have long since wound themselves up to a much greater degree than we observe; they are thought to consist of density waves that drift through the stars and gas like ripples on a pond. The spiral arms are delineated by their overabundance of bright, young stars and glowing gas clouds, and may

Types of galaxies



Fig. 4.17 The giant elliptical galaxy M87, located at the heart of the Virgo Cluster. An elliptical galaxy is an ellipsoidal mass of stars, showing no overall rotation or structures such as spiral arms. (Image copyright AURA Inc./NOAO/NSF.)

be the major location for star formation. Spirals have a range of sizes, containing from a few billion to several hundred billion stars.

The other major category of galaxy is the *elliptical galaxies*. As their name implies, these galaxies are ellipsoidal, that is, shaped roughly like a football. Some are, or appear to be, nearly spherical, especially the largest ones. Ellipticals cover an enormous range, from the dwarfs, with as few as a million stars, up to the giants, which contain thousands of billions of stars. In contrast to spirals, ellipticals seem to contain scant gas or dust and show little evidence of recent star formation. The third category is something of a catch-all for any galaxy that does not fit into the previous two: the *irregular galaxies*. Irregular galaxies show no particular structure, though many might be distorted by their interactions with other galaxies. Some irregulars, especially the *dwarf irregulars*, might be prevented from pulling themselves into a spiral shape by the gravitational dominance of large galaxies that they orbit. Others may simply show no structure, or even a tendency toward structure, at all. How galaxies formed, why they take the shapes they do, and why so few types are observed, are some of the major outstanding puzzles of cosmology and astronomy.

The galaxy in which the Sun and its solar system are located is called the Milky Way Galaxy, or just the Galaxy. Though we cannot, of course, observe it from the outside, the distribution of stars in our skies immediately shows that the Milky Way consists of a flat disk. We cannot see its center in visible light because thick clouds of obscuring dust intervene between us and the core, but we know that the center of the Milky Way lies in the constellation Sagittarius and is one of the brightest radio sources in the Galaxy. Our inability to see our own Galaxy from the exterior inhibits detailed understanding of its structure. We can, however, estimate it to contain approximately 100 billion stars. The Sun is about 30,000 lightyears from the center, roughly two-thirds of the way to the visible edge of the Galaxy. (Galaxies have no strict cutoff, but at some point become faint enough to define a boundary.) The solar sys-

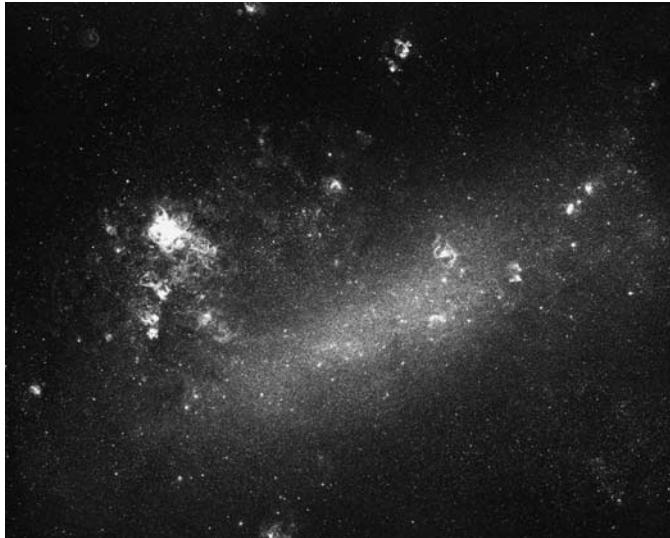


Fig. 4.18 The Large Magellanic Cloud is an example of an irregular galaxy. This small galaxy is a satellite of the Milky Way and can be seen from the Southern Hemisphere of Earth. (Image copyright AURA Inc./NOAO/NSF.)

tem completes one revolution around the Galactic center in 200 million years.

Galaxies show a strong tendency to bunch into **galaxy clusters**. Our own Local Group is a modest cluster of perhaps a few dozen galaxies, dominated by the two large spirals called the Milky Way and the Andromeda Galaxy. This is a typical configuration for such loose clusters and, like others observed, ours is asymmetrical; the Andromeda Galaxy is about twice as massive as the Milky Way. Other galaxy clusters are much richer and denser, containing anywhere from a few hundred up to thousands of large galaxies, and unknown numbers of small, dim galaxies. Whereas the dominant galaxies of loose clusters are generally spirals, rich clusters contain a mixture of galaxy types. Most ellipticals reside in fairly dense clusters, and giant ellipticals are often found at the very center of a large cluster. The spatial scale of galaxy clusters also varies, from the 2 million lightyears of the Local Group to the 6 million lightyears of a rich cluster such as the Virgo Cluster. Galaxy clusters are gravitationally bound; that is, the galaxies orbit one another. Beyond this scale is the suggestion of even larger structures, called **superclusters**. The largest superclusters seem to be too large to be fully gravitationally bound; their origin is a mystery. Perhaps the galaxies are merely particles in some great overarching structure of the universe. How big are the largest structures, and how could they have originated? These are some of the most important questions in cosmology.

We do not know how many galaxies inhabit the universe. Beyond the reaches of the Milky Way itself, nearly every object we can see is a galaxy. There are at least as many galaxies in the universe as there are stars in the Milky Way. Galaxies may be the glowing tracers of the mass of the universe, the visible spots in a great flow of matter; or they may contain most or all the matter. Galaxies formed very early in the history of

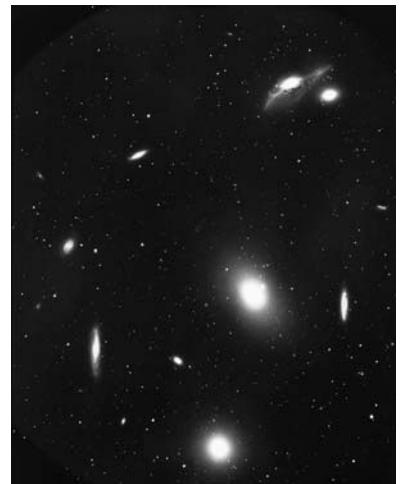


Fig. 4.19 The center of the great galaxy cluster in the constellation Virgo. This irregular cluster contains approximately 2,500 galaxies. (Image copyright AURA Inc./NOAO/NSF.)

The largest structures in the universe are superclusters of galaxies

the universe; but are galaxies fundamental, or did they condense from the larger structures we observe? What creates and maintains spiral patterns? Do ellipticals result from the merger of spirals, or are they of different origin? Galaxies are the starry messengers which tell us of the origin and structure of the universe itself, if only we could understand their stories.

Chapter Summary

As chemistry and physics began to develop into their modern forms in the late 18th and early 19th centuries, scientists began to elucidate the nature of matter. The elements were chemically irreducible substances; it was soon realized that they corresponded to unique types of atoms. The discovery of elementary particles clarified chemical behavior; the number of protons defined the elemental identification of the atom, while the electrons accounted for its chemical properties. The structure of atoms was mysterious until experiments by Ernest Rutherford proved that they consist of a compact nucleus surrounded by a cloud of orbiting electrons. The discovery of the neutron ushered in the era of nuclear physics. The number of protons in the nucleus specifies the element, while the number of neutrons fixes the isotope. Nuclear reactions occur for certain isotopes that can change the number of protons and neutrons in the nucleus, thus altering the identity of the atom. Radioactivity can cause the atom to shift several places on the Periodic Table. Even more extreme nuclear reactions can occur; fission splits a heavy atom into two or more daughter atoms, while fusion joins two or more light atoms into a heavier element.

A full understanding of the atom required the development of the theory of quantum mechanics. Observables such as position, velocity, energy, and so forth are not infinitely variable, but are multiples of irreducible quantities called *quanta*. The spin of a particle can quantize to a full-integer multiple of a base quantum or to a half-integer multiple; this divides the elementary particles into two families, the bosons and fermions respectively. Bosons carry forces, while fermions make up matter. Fermions obey the exclusion principle; two interacting fermions cannot occupy the same overall quantum state. Quantum mechanics was combined with special relativity to predict the existence of antimatter, which was confirmed with the discovery of the positron.

Quantum mechanics led to a greater understanding of the forces of nature. All forces result from one of the

four fundamental forces. The strong and weak interactions operate at nuclear length-scales and govern nuclear reactions; the strong force also holds protons in nuclei together against their mutual electrostatic repulsion. The electromagnetic force, which has an infinite range, occurs between charged particles and accounts for the majority of everyday forces. The gravitational force occurs between massive objects; it also has infinite range and, since it is never shielded, or partially cancelled due to opposite charges, it governs the large-scale interactions of the universe and its contents.

Work occurs when a force is exerted to produce a displacement. Energy is that quantity which represents the capacity to do work. Energy comes in many forms: chemical energy, potential energy (energy due to location in a potential field), kinetic energy (energy of motion), energy of deformation, and so forth. An important form of energy is heat, which is the aggregate energy due to random motions of the particles that make up a substance. The law of conservation of energy states that energy is never created or destroyed, but is only converted from one form into another. In nonrelativistic physics, a companion law, the law of conservation of mass, states that mass is not created or destroyed. Thermodynamics is the science of energy in general and heat in particular. The second law of thermodynamics is one of the fundamental laws of the universe; this law states that in a closed system the total entropy, which is a measure of disorder, at best remains the same and in general always increases.

Complementary to matter is the world of waves. A wave is a propagating disturbance in some quantity. A pure wave has a frequency and a wavelength and its strength is given by its amplitude. When multiple waves of different frequencies combine, the result is a spectrum. Waves have unique behaviors; they can refract, reflect, and diffract. When the source or receiver of a wave are moving relative to one another, the wavelength and hence the frequency of the wave is shifted; this is the Doppler effect. If source and receiver are approaching, the wave-

length decreases; this is a blueshift. Conversely, if the source and receiver are receding relative to one another, the wavelength increases; this is a redshift. Wavelength shifts due to the Doppler effect provide much of the information available to astronomers from distant objects.

Radiation is the emission of energy from an object, often in the form of a wave. Every object with a temperature above absolute zero emits thermal radiation; the higher the temperature, the more energetic the radiation. In general, the spectrum of thermal radiation is a complicated function of the temperature, composition, shape, and other properties of the object, but in one special case, that of blackbody radiation, the spectrum depends only upon the temperature of the emitter. A blackbody is a perfect absorber (and hence also a perfect emitter); blackbody radiation represents a state of equilibrium in the emission. Many real objects can be approximated as blackbodies, making blackbody radiation useful despite its requirement of ideal conditions. Moreover, the study of blackbody radiation led to the creation of quantum mechanics.

The total quantity of radiated energy per unit time is the luminosity of the object. Light from luminous ob-

jects is nearly our only source of information about the universe outside our solar system. As the energy travels into space, it is spread out over a sphere of ever-increasing diameter. Consequently the brightness of the emitter diminishes inversely to the square of the distance. If all other sources of energy loss can be ignored, the observed brightness of an object of known luminosity would immediately give its distance. Distance computed in this manner is called the luminosity distance.

The most prominent inhabitants of the universe are the galaxies. There are at least as many galaxies in the observable universe as there are stars in the Milky Way Galaxy. Galaxies are categorized by their structure as spiral, elliptical, or irregular. Galaxies show a strong tendency to bunch into galaxy clusters, gravitationally bound agglomerations of galaxies whose size ranges from a few dozen members, such as is the case for our own Local Group, to enormous rich clusters containing thousands of members. Beyond the clusters is evidence of even larger structures, the superclusters. The mystery of the origins of the largest superclusters remains one of the most important questions in cosmology.

Key Term Definitions

photon A boson which is the particle of electromagnetic radiation (light). The photon is also the carrier particle of the electromagnetic force.

nucleon Either of the two fermionic particles, the proton and the neutron, which form the nuclei of atoms.

isotope One of the forms in which an element occurs. One isotope differs from another by having a different number of neutrons in its nucleus. The number of protons determines the elemental identity of an atom, but the total number of nucleons affects properties such as radioactivity or stability, the types of nuclear reactions, if any, in which the isotope will participate, and so forth.

nuclear reaction A reaction that occurs in and may change the nucleus of at least one atom. Examples include radioactivity, fission, and fusion.

neutrino Any of three species of very weakly interacting lepton with an extremely small mass.

boson A class of elementary particles whose spin is an integer multiple of a fundamental quantized value.

The major function of bosons is to mediate the fundamental forces. The best-known boson is the photon.

fermion A class of elementary particles whose spin is a half-integer multiple of a fundamental quantized value. Fermions make up matter. The best-known fermions are protons, neutrons, electrons, and neutrinos. Fermions obey the exclusion principle.

exclusion principle The property that fermions of the same type that are able to interact with each other cannot simultaneously occupy the same quantum state.

antimatter Particles with certain properties opposite to those of matter. Each matter particle has a corresponding *antiparticle*. The antiparticle has exactly the same mass and electric charge as its partner. When a particle combines with its antiparticle both are annihilated and converted into photons.

strong interaction The fundamental force that binds quarks into hadrons and holds nucleons together in

atomic nuclei. Sometimes called the strong force or the strong nuclear force.

weak interaction The fundamental force that accounts for some particle interactions, such as beta decay, the decay of free neutrons, neutrino interactions, and so forth. Sometimes called the weak force or the weak nuclear force.

electromagnetic force The force between charged particles that accounts for electricity and magnetism. One of the four fundamental forces of nature, it is carried by photons and is responsible for all observed macroscopic forces except for gravitational forces.

gravity The weakest of the four fundamental forces; that force which creates the mutual attraction of masses.

energy The capacity to perform work, where *work* is defined as the exertion of a force to produce a displacement.

potential energy The energy possessed by something by virtue of its location in a potential field, for example, its position in a gravitational field.

kinetic energy The energy associated with macroscopic motion. In Newtonian mechanics, the kinetic energy is equal to $\frac{1}{2}mv^2$.

thermodynamics The theory of heat and its relationship to other forms of energy.

heat A form of energy related to the random motions of the particles (atoms, molecules, etc.) that make up an object.

temperature A measure of the average kinetic energy of random motion of the constituents (for example, molecules, atoms, or photons) of a system.

conservation of energy The principle that the total energy of a closed system never changes, but energy is only converted from one form to another. This principle must be enlarged under special relativity to include mass-energy.

conservation of matter The principle that matter is neither created nor destroyed. This principle is

only approximately true, since it does not hold in special relativity.

entropy A quantitative measure of the disorder of a system. The greater the disorder, the higher the entropy.

spectrum The components of emitted radiation, or a collection of waves separated and arranged in the order of some varying characteristic such as wavelength, frequency, mass, or energy.

interference The interaction of two waves in which their amplitudes are reinforced and/or cancelled.

Doppler effect The change in frequency of a wave (light, sound, etc.) due to the relative motion of source and receiver.

blueshift A shift in the frequency of a photon toward higher energy.

redshift A shift in the frequency of a photon toward lower energy.

lightyear (ly) A measure of distance equal to that traveled by light in one year.

blackbody radiation A special case of thermal radiation, emitted by a blackbody and characterized by thermal equilibrium of the photons. A blackbody spectrum is completely determined by the temperature of the emitter.

equilibrium A balance in the rates of opposing processes, such as emission and absorption of photons, creation and destruction of matter, etc.

luminosity The total power output of an object in the form of light. (Sometimes extended to include the output of all forms of radiated energy.)

galaxy A large, gravitationally bound system of stars, star clusters, and interstellar matter.

galaxy cluster A group of galaxies that are mutually gravitationally bound.

supercluster A cluster of galaxy clusters.

Review Questions

- (4.1) What is an isotope, and how is it related to an element? Why does atomic number determine the chemical properties of an element?
- (4.2) ^{238}U (uranium) decays to ^{206}Pb (lead) with a half-life of 4.5 billion years. If the ratio of $^{238}\text{U}/^{206}\text{Pb}$ in a meteor is equal to 1/3, how old is the sample? If the meteor originally contained some ^{206}Pb from a source other than radioactive decay, how would that affect your age estimate?
- (4.3) Describe and distinguish nuclear fission and fusion. Which types of element are involved in each of these processes?
- (4.4) Describe two differences between a boson and a fermion. To which family does the electron, the proton, and the photon belong?
- (4.5) Name the four fundamental forces of nature. Which are the strongest? Which one creates most everyday forces? Which one dominates the universe at large scales? Why does only one force dominate at large scales?
- (4.6) The wavelength of a particular hue of green light is 5.0×10^{-7} meter. What is the frequency of this light? What is the energy of a photon of this light? (Values of some important constants of physics and astronomy are given in Appendix C.)
- (4.7) A man comes before a traffic court, charged with going through a red light. He argues that the Doppler shift made the light appear green to him. If red has a wavelength of 7000 Ångstroms (Å; one angstrom is equal to 10^{-8} centimeters) and green has a wavelength of 5500 Å, then, from the Doppler shift formula, what was his speed as a fraction of the speed of light c ?
- (4.8) The diameter of a telescope's mirror determines how much light it can gather. The amount of energy collected over the area of the mirror from the light of a particular star can be measured with sensitive instruments. How does the inverse square law then tell you the *total* energy given off by that star? At the Earth's orbit the light of the Sun is distributed evenly over a sphere with a radius equal to that of the Earth's orbit (about 10^{11} m.). The telescope has a 1-meter radius (2-meter diameter). What fraction of the Sun's light can the telescope capture?
- (4.9) What is the unique characteristic of blackbody radiation?
- (4.10) How does the surface temperature of a reddish star compare with the surface temperature of a bluish-white star? Does the diameter of the star matter when determining the temperature? What is the ratio of the peak wavelength emitted by Star 1 to the peak wavelength emitted by Star 2 if the surface temperature of Star 1 is twice that of Star 2?
- (4.11) Explain the significance of luminosity distance. What sort of errors can occur in the measurement of this quantity? How can astronomers correct for these complications?
- (4.12) The Andromeda Galaxy is about 2 million lightyears away from us. To what distance would that correspond in the scale model discussed in this chapter, in which the Sun is the size of an orange?

This page intentionally left blank

The Lives of the Stars

5

...the glorious sun
Stays in his course, and plays the
alchemist

William Shakespeare, *King John*

The stars change very little during the course of a human lifetime. Indeed, they have hardly changed in appearance over the length of recorded human history. In pre-Newtonian cosmologies, the stars were eternally affixed to a single, unchanging celestial sphere. Even after the age of Newton, they were the fixed stars, whose distribution coincided with absolute space. The occasional appearance of a supernova indicated that perhaps the heavens were not immutable, but it was only in the 20th century that these rare events were associated with the deaths of stars. As for stellar birth, astronomy textbooks dated as recently as the 1950s speculate that there might perhaps be places where we could observe a new star being formed, as if such an event would be quite rare.

We know now that stars are not eternal; they come into existence, go through a life cycle, and die. Through observations, and through the careful construction of detailed models based on an understanding of the laws of physics, astronomers have learned a good deal about the lives of the stars. The type of existence a given star has, and the circumstances of its death, depend upon the *mass* of the star, and to a lesser extent upon its *chemical composition*. Less massive stars, such as the Sun, burn their fuel slowly and live long; when they exhaust their fuel stores, they flicker out as slowly cooling white dwarfs. More massive stars live fast and die young, and end their existences in some spectacular cataclysm, leaving behind a compact and enormously dense cinder called a neutron star. The most massive stars have the most violent ends; they may blow themselves to nothingness in a supernova or, if a core is left behind, they may collapse until they cut themselves off from the rest of the universe.

In comparison to the grand galaxies that fill the huge volume of the universe, individual stars might seem insignificant. Certainly, one smallish star is of great importance to life on one tiny planet, but what roles might stars play in the cosmos at large? Most obviously, the stars make it possible for us to be aware that anything else exists. If all matter other than the Sun were dark, we would not even know, at least directly, of our own Galaxy, much less of the billions of other galaxies that fill the universe. Some light is emitted from very hot gas near the centers of galaxies, but most of the visible light in the universe, and much of the energy in other bands, originates directly or indirectly from stars. The

Key Terms:

- interstellar medium
- nebula
- brown dwarf
- hydrostatic equilibrium
- ideal gas
- deuterium
- primordial element
- main sequence
- Population I, II, III
- metal
- globular cluster
- turnoff mass
- red giant
- electron degeneracy
- white dwarf
- accretion disk
- nova
- Chandrasekhar limit
- supernova
- neutron degeneracy
- neutron star
- pulsar
- conservation of angular momentum

Stars are a fundamental component of the cosmos

populations of stars make galaxies visible, but more than that, they enable us to measure the masses and compositions of the galaxies. Certain kinds of bright stars provide a means to gauge the distances of galaxies; furthermore, when a massive star collapses, the resulting explosion is so brilliant that it can be seen across enormous expanses, providing a means to measure the distance that the light has crossed. Humbler, lower-mass stars have an equally important role to play in our cosmological investigations. Such stars can have ages comparable to the age of the universe itself. A star is a much simpler object, and much more amenable to observation, than is the universe as a whole, so that stars provide us with an independent estimate for the age of the cosmos. Finally, stars play an active role in the evolving cosmos; their nuclear furnaces are the sole source of all the elements beyond lithium. As arguably the most important denizens of the universe, the stars are of great importance in the study of cosmology.

A star is born

Interstellar gas is the raw material from which stars form

Between the stars lies interstellar space. By Earthly standards it is an excellent vacuum, with an average density of about one atom of hydrogen per cubic centimeter. Nevertheless, within this space are enormous clouds of gas, consisting mostly of hydrogen with a lesser quantity of helium. Some clouds also contain considerable *cosmic dust*, which consists primarily of tiny specks of minerals and soot, sometimes coated with various ices. The matter between stars is collectively known as the **interstellar medium**. Clouds of interstellar material, although very tenuous, are so large that their masses can be quite significant, up to thousands of solar masses. They are called **nebulae** (singular *nebula*), from the Latin word for cloud. One important effect of the nebulae is their influence on starlight. They absorb some of the photons and scatter, that is, send in all directions, others. Looking at stars through interstellar clouds is much like trying to see the headlights of vehicles through a fog. How much light is transmitted depends upon the thickness of the fog, as well as on the types of particles that make it up. Some nebulae, mainly those that are very cold and contain much dust, are almost completely opaque. Other nebulae contain bright stars embedded within them, and glow themselves due to their re-emission of the energy they absorb from the star. In any case, their presence complicates our measurements of the luminosity distances to stars that are partially obscured by them. The resultant dimming of the stars' light makes them seem farther than they really are.

But obscuration is far from the only role these great clouds play in the universe; their most important function is to be the birthplace of stars. A star is born when a cloud of gas contracts under its own gravity. Today such a statement may seem obvious, but it was a daring hypothesis when it was first put forward late in the 18th century. The philosopher Immanuel Kant, among others, had arrived at inklings of

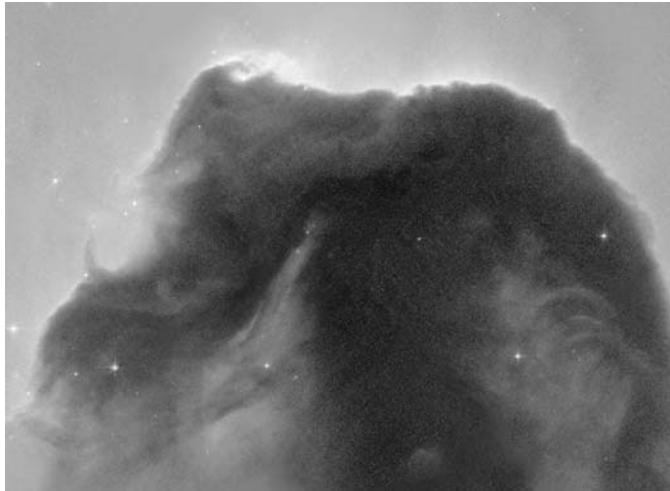


Fig. 5.1 The Horsehead Nebula, a dark cloud of gas and dust silhouetted against glowing gas in the constellation Orion. The bright gas visible at the top left edge has been heated by a young star still embedded in the dark nebula. (NASA/NOAO/ESA, and the Hubble Heritage Team.)

this model, but it was first introduced in a developed form by Pierre Simon de Laplace. Laplace proposed that a rotating cloud of gas would flatten into a disk as it pulled itself together. The central portion of the disk would gather itself into a ball to form a star, while the outlying regions serenely coalesced into planets. The disk hypothesis neatly explained why all the planets known to Laplace obediently orbit nearly in a single plane. Regardless of how appealing the picture, however, for it to be a scientific model it must be developed quantitatively. During his lifetime the mathematical tools barely existed to study his proposal carefully; indeed, Laplace himself invented many analytical techniques for working with gravitating systems. It should not be surprising, then, that the details of his model did not quite work out. Nevertheless, Laplace's insight is still a useful conceptualization today, and it gives a good qualitative description of the process that creates new stars.

Star formation is still understood only in outline; the details remain elusive and are a subject of active research. The basic ideas are simple, however. The Galaxy is filled today with clouds of gas, as must have been even more true in the past. The most likely stellar nurseries are gigantic *molecular clouds*, huge aggregations of cold gas, mainly hydrogen. Whenever possible, hydrogen forms a molecule consisting of two atoms. In the near vacuum of interstellar space a lone hydrogen atom has little opportunity to encounter another one, and most of the hydrogen is atomic. A molecular cloud, however, has a sufficient density that about half its gas takes the molecular form. In the present-day Galaxy, these clouds generally also contain many other kinds of relatively simple molecules such as carbon monoxide, water, formaldehyde, ethanol, and ammonia. The molecular clouds also have an abundance of dust grains. Dust grains are very efficient at radiating away energy, which is an important reason that these clouds are likely progenitors of stars; the dust cools the cloud and helps to shield the molecules from high-energy photons, leaving the clouds with temperatures ranging from 10 K to

Molecular clouds are the birthplaces of stars

Stars form when gravity overcomes pressure

100 K. Dust is also particularly opaque to most visible light, scattering it away so that it never reaches our telescopes; thus clouds containing dust grains are dark, and the dustier the cloud, the blacker it appears through a telescope.

Gravity will, as always, try to pull dense regions into collapse; the cloud resists this through gas pressure. Variations in pressure in a fluid are transmitted by means of sound waves. If sound waves can cross the condensing region faster than it contracts, then the waves will be able to restore a balance between gravity and the pressure of the gas. The speed of sound decreases with temperature; therefore, the colder the gas, the greater its chance of collapsing before pressure can build up. Under the right conditions the cloud, or a portion of it, will be unable to maintain itself and will begin to contract. The same gravitational instability that initiates the contraction will probably also cause the condensing region to fragment into many smaller clumps; thus most stars likely form as members of clusters. There are many known clusters of young stars, such as the famous Pleiades, and most stars that are known to be young reside in groups. Many older stars, however, travel solo through space, or perhaps in the company of one or a few other stars. An isolated star, such as the Sun, probably escaped in its youth from its nursery mates due to gravitational interactions among the young stars, and with external objects.

Gravitational contraction produces heat

As each would-be star collapses due to its self-gravity, the gas retains its spin, or angular momentum, and forms a disk much like that imagined by Laplace. The collapse compresses the gas, causing it to heat up. Some of the rotation of the gas is carried away by magnetic fields threading the cloud, allowing further collapse and compression at the center. Eventually most of the matter accumulates at the center, while the rest remains in an encircling disk. The central sphere, now a protostar, continues to contract and heat. As its temperature rises, more and more of its hydrogen ionizes, that is, the atom loses its single electron. Free electrons scatter and absorb photons very effectively, so the more electrons that are liberated, the more opaque the protostar becomes. If photons cannot escape from the gas, their energy is trapped within the protostar, causing the temperature to rise even further. If the temperature within the core rises to a sufficient level to ignite nuclear fusion, the energy generated from this process provides the newborn star with the pressure required to prevent further collapse.

Meanwhile, the protostellar disk is undergoing changes of its own. The heat and pressure from the particles and photons streaming from the surface of the infant star blow the lightest and most volatile elements, such as hydrogen and helium, away from a region immediately surrounding the star. Hence the clumps, or *planetesimals*, that are able to form near the star are likely to be rocky, composed of mostly nonvolatile substances. Farther away, planetesimals form with large quantities of lighter matter such as ices. As the young stellar system develops, these planetesimals collide and clump together; the largest clumps sweep up the smaller particles they encounter, becoming larger *planetoids* in the

process. Planetoids in the outer, cooler part of the disk can attract and retain hydrogen and helium, becoming gas giant planets. A new solar system has formed.

To create a star, the core temperature in the collapsing protostar must rise high enough to ignite nuclear reactions. It is likely that many globules of gas that begin to contract are too small for this ever to occur. Calculations show that the minimum mass for star formation is in the neighborhood of $0.08 M_{\odot}$.¹ Condensed objects below this mass limit cannot produce a sufficiently high temperature at their cores to initiate nuclear fusion. The fraction of stars that form with a given mass seems to be mostly determined by the mechanisms of star formation; when we consider all such fractions for all masses, we obtain a function called the *initial mass function* (IMF). Unfortunately, the IMF is only partially known; the uncertainties for low-mass stars are substantial. The observations suggest that there may be differences in the star-formation process for low- and high-mass stars, with the division point at about a solar mass. The IMF also seems to indicate a diminished efficiency of star formation for masses below approximately $0.2 M_{\odot}$, which is considerably greater than the theoretical minimum mass for a star. It may be that some factor other than mass alone restricts the number of stars at the lowest mass range.

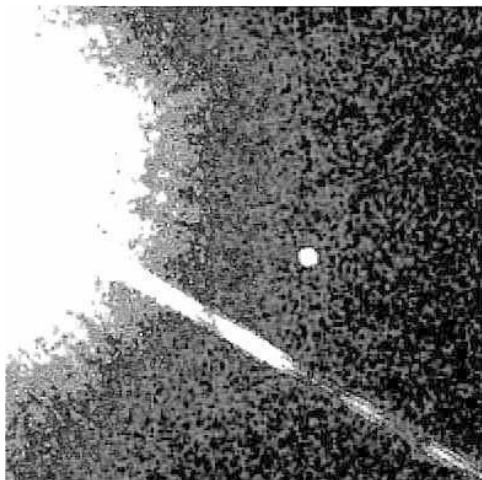
Humble objects are the most abundant in nature, and this is a simple fact, not an anthropomorphism; massive stars are rare, while the majority of stars are smaller than the Sun. At the low end of the mass range the stars are cool, glowing only with a faint red light; these stars are called *red dwarfs*. Those objects that are just below the mass required for stellar ignition have been dubbed **brown dwarfs**. But just how many such failed stars exist? They might be scattered throughout the Galaxy and could, if they exist in large numbers, make a significant contribution to the total mass of the Galaxy. Brown dwarfs might still radiate heat as they slowly contract, converting some of their gravitational energy into infrared radiation; in principle, this infrared signal could be detected. They are extremely dim even in the infrared, however, making them very difficult to see. Nevertheless, new technology has made it possible to look for brown dwarfs, and the dimmest of the red dwarfs. Deeper searches have found evidence for very dim red dwarfs, although not in the numbers originally expected. One of the first candidates for a brown dwarf was found late in 1995; it is the tiny companion to a star known as Gliese 229. More recent surveys have found more brown dwarfs; their numbers, however, are such as to make up no more than about one quarter the total number of stars in the Milky Way. Moreover, because each brown dwarf's mass is so low, their total makes an insignificant contribution to the mass of the Milky Way. It seems likely that this conclusion is general, and that brown dwarfs are not an important component in the total mass density of the universe.

Nuclear reactions in the core define a star

Low-massed stars are most abundant

¹The symbol M_{\odot} represents the mass of the Sun.

Fig. 5.2 *Hubble Space Telescope* image of the faint brown dwarf companion of the cool red star Gliese 229. The brown dwarf is located 19 lightyears from Earth in the constellation Lepus. Estimated to be 20–50 times the mass of Jupiter, it is too massive and hot to be a planet but too small and cool for a star. (T. Nakajima and S. Kulkarni, Caltech; S. Durrance and D. Golimowski, JHU; STScI/NASA.)



Stellar systems

We have given only the barest sketch of the formation of stars; there are sure to be many variations on this theme. Many effects are still poorly understood, such as the role that might be played by magnetic fields. Moreover, our qualitative description might seem to apply only to solitary stars; yet many, perhaps half or more, of all stars are members of binary systems, two stars that orbit one another. Some stellar systems of three or even four stars exist nearby. For example, Mizar, a star in the handle of the Big Dipper (Ursa Major), is a doublet; those with good eyes can easily make out the companion, Mizar B, on a dark night. Many such *optical doubles* are coincidental, the two stars being at vastly different distances, but Mizar A and B are, in fact, a pair; they comprise a *visual binary*. It turns out that both Mizar and Mizar B are themselves double stars, making the system a quadruple star! Yet the broad outline we have sketched surely still applies to such stars. Whether each star within a system might have an associated disk, at least near the time of its birth, is uncertain. The disks may be disrupted if the companion is too close, or they might survive but be unable to produce planets, or there may be planets around the members of some binary systems. It is difficult enough to understand thoroughly the formation of one star and its disk; multiple-star systems are another step upward in complexity.

Astronomers cannot even be certain that a planetary system, or even a protoplanetary disk, forms around all single stars. Theory indicates that it should, although the subsequent formation of planets may well not be inevitable. Even if they form, they may not survive; the wind of high-energy particles and the intense radiation from a very massive, bright star might sweep away all of its disk, not just carve out a small region depleted in light elements. On the other hand, there is direct evidence that the solar system is not unique in the Galaxy, much less in the universe. One nearby, young star, Beta Pictoris, has yielded photographs of a disk of dust. Disks are difficult to detect even for close stars; the glare from Beta Pictoris ordinarily overwhelms the weak emissions from

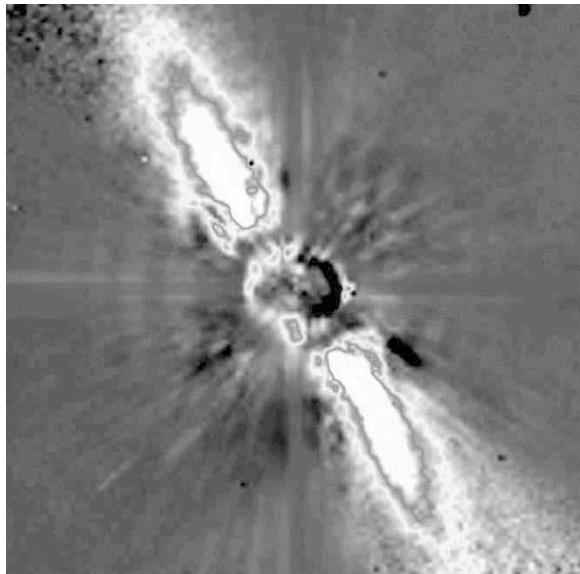


Fig. 5.3 Beta Pictoris, a young star in the constellation Pictor, is surrounded by a disk of gas and dust that may be the progenitor of a planetary system. (Copyright ESA/ESO; prepared by G. Blake, Caltech.)

its disk, most of which are in the infrared, and the star's image must be artificially covered for the disk to become visible. Unfortunately, no planets at this distance could possibly be directly resolved in the disk, even if they might be present or forming. Hence all we can conclude with certainty from Beta Pictoris is that it provides a wonderful example of a star that does possess a dusty disk during the early stages of its life.

Astronomers have gone even further in their observations of stellar systems. Late in 1995, the first extrasolar planet orbiting an ordinary star was found around 51 Pegasus, a star similar to the Sun. Discovered by astronomers at the Geneva Observatory and confirmed by observations at the Lick Observatory, the planet has at least half the mass of Jupiter. This planet cannot be seen directly even with powerful telescopes; its presence was inferred from the wobble in the star's motion produced by the planet's gravity. Since then, astronomers have detected many additional planets using this technique. By now well over a 100 of these extrasolar planets have been discovered, all associated with stars that are near to the Sun. These findings suggest that planetary systems must be relatively common. All of the planets discovered so far have been large, comparable in mass to Jupiter; a planet as small as the Earth would be extremely difficult to detect even with improving technology, though eventually it may become possible. In any case, there is now unequivocal proof that some other stars do have planetary companions, though as yet no firm evidence exists of extrasolar planets suitable for the formation of life as we know it. Most astronomers were always confident that other planetary systems would be found, though they were still quite excited by these discoveries; at last, there is more than one such system to study, so that theories of planet formation may begin to be tested.

The discovery of extrasolar planets

Holding its own

Stars remain in a balance between inward-directed gravity and outward-directed pressure

What, then, is a star? All stars are huge balls of gas, mostly hydrogen, held together by gravity. Throughout the life of a star, two opposing forces determine its structure: gravity and pressure. Gravity works to pull the gas toward the center of the star; as the gas is compressed under its own weight, its pressure rises until a balance is reached. This state of balance between two competing forces is known as **hydrostatic equilibrium**, and it holds for most of the lifetime of a star. To understand stars, we must understand how they generate and radiate the energy that offsets the omnipresent pull of gravity. This much was long understood, but at the beginning of the 20th century the mystery was the mechanism of energy generation. One possibility, ordinary chemical reactions, is certainly insufficient to keep the stars burning for very long. If the Sun were made entirely of coal, and some source of oxygen allowed the coal to burn, a star's entire life would last only a few hundred thousand years. But geologists had plenty of evidence that the Earth was much older than this. In the late 19th century, the physicists Hermann Helmholtz and Lord Kelvin (William Thomson) independently suggested an alternative power source: gravity itself. Energy is released when a body is dropped in a gravitational field; for example, water falling over a water wheel performs work. Perhaps, Kelvin and Helmholtz conjectured, the balance provided by hydrostatic equilibrium was not quite perfect. Perhaps the star continued to contract under its own gravity at a very slow rate. As it did so, its gas would be compressed and heated, so that some of its gravitational energy would be converted into heat and light.

The belief that gravity powered the stars held sway for many years, although there were hints that this was not correct. Calculations indicated that gravity could keep the Sun shining for many millions of years, but mounting terrestrial evidence suggested that the age of the Earth was in the *billions* of years. The discovery of radioactivity near the end of the 19th century provided a possible solution to this conundrum. Here was a previously unknown energy source, clearly neither chemical nor gravitational. As more and more came to be known about the atom, physicists realized that the nuclei of atoms could be broken apart or fused together and that in many cases this would release energy, possibly in enormous quantities. Einstein's famous formula $E = mc^2$, which in essence states that energy and mass are equivalent, shows just how much energy nuclear reactions can release. Multiplying the mass of the Sun by the speed of light squared, c^2 , and dividing by the solar luminosity, the energy radiated per second, shows that the upper limit for the Sun's lifetime would be

$$\frac{M_\odot c^2}{L_\odot} = 14,000,000,000,000 \text{ years.} \quad (5.1)$$

Nuclear reactions make possible the long lives of the stars

Thus only a small percentage of the total mass of the Sun need be converted to energy to enable it to burn for tens of billions of years. Nuclear reactions could easily provide more than enough time for the

Earth and its inhabitants to form and evolve. It was first suggested in the 1920s that protons might fuse to form helium in the cores of stars, liberating energy. However, further progress in nuclear theory was required before scientists had the definitive answers. By 1938, enough was known about nuclear physics for physicists to work out the details of one particular sequence of fusion reactions by which the stars shine.

The stars we see in the sky, and our own Sun, are furnaces burning nuclear fuel. The heat generated by those nuclear reactions provides the gas pressure to keep the star from collapsing under its own weight. How can a gas accomplish such a Herculean task? The gases inside the Sun consist of the nuclei of atoms, and the electrons that have been stripped from those nuclei, all moving about at high speed and colliding with one other. Is there any way to make sense of this chaos? Fortunately, there is. The number of particles is so huge, and the way they interact sufficiently simple, that the behavior of the gas as a whole can be described in an averaged, statistical way. Any particular particle will have some mass m , and will be moving at some velocity v , until it collides with some other particle and changes its velocity. Velocity and particle mass can be combined to yield an energy due to motion, that is, a *kinetic* energy. The Newtonian formula for this energy is $E_k = \frac{1}{2}mv^2$; this holds for any particle as long as Newtonian physics is valid, which is mostly true even in the interior of the Sun. Because the collection of particles is constantly interacting, the gas comes into an equilibrium characterized by some average particle kinetic energy. The quantity we call *temperature* is defined by this average energy per particle. The higher the average kinetic energy of the particles, the higher the temperature. Moreover, two gases that have the same temperature but different particle masses must differ in their average particle velocities; the gas with the lower-mass particles would necessarily have a higher average velocity.

If a gas at temperature T is confined within a rigid box, the particles will collide not only with each other, but also with the walls of the box. Since each collision changes the velocity of the incident particle, a force must be exerted upon the particle; but by Newton's third law, the particle must also exert a force upon the wall. Multitudes of such collisions by the constituent particles of the gas can be averaged to yield a macroscopic force per unit area upon the wall, resulting in *gas pressure*. Working through the units, we find that force per unit area has the same dimensions as energy per unit volume; indeed, pressure can also be characterized by the average energy in a gas per unit volume. But we have just argued that for a gas in equilibrium, the temperature specifies the average kinetic energy per particle. A higher average kinetic energy ($\frac{1}{2}mv^2$) should result in a larger change in momentum (mv) when gas particles strike their surroundings. We conclude that temperature and pressure ought to be related. For many gases, including stellar gas, the pressure and temperature are related by the **ideal gas law**

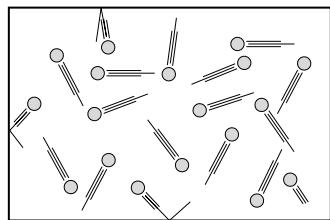


Fig. 5.4 Gas particles move about with an average speed that increases with the temperature of the gas. Collisions with the walls of the box exert a force per unit area, or pressure, on those walls.

Pressure from an ideal gas

$$P = nk_B T. \quad (5.2)$$

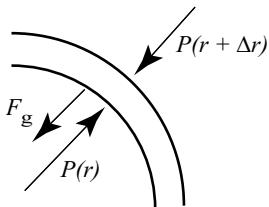


Fig. 5.5 Forces on a spherical shell of gas in a star. The gravitational attraction of the mass interior to the shell creates a downward force. The pressure from gas above the shell pushes downward, while the pressure from gas below exerts an upward force. The upward pressure force is greater than the downward, and this difference balances the gravitational force.

Elements of stellar structure theory

This equation states that pressure P equals the number of gas particles per unit volume n , multiplied by the average energy, $k_B T$, of those particles. The quantity k_B is called the *Boltzmann constant*; it provides the connection between temperature and energy. The temperature T must be measured on a scale that sets its zero point at absolute zero. The ideal gas law shows that the higher the temperature, the more rapid the motions, the larger the kinetic energies, and the greater the pressure; specifically, the temperature and pressure are proportional for a fixed amount of gas held to a fixed volume. Of course, a star is not a rigid container; its gas has some freedom to expand or contract, and such changes in volume affect the pressure and temperature through the number density n . However, under most circumstances during a star's life in equilibrium, the changes in volume are quite small relative to the size of the star.

A star must be supported by the pressure of its gas. The deeper into a star, the greater the weight of the overlying layers, and the higher the pressure and temperature must be. Consider a thin shell of gas located at a radius r from the center of the star. If the star is to remain stable, the net inward force on any such shell must equal the net force outward. The net force inward is the force of gravity at that location, plus the inward-directed force due to the pressure from the gas lying beyond r . The only available outward force is provided by pressure from the gas beneath the shell. Setting the outward force equal to the total inward force leads to the conclusion that pressure must increase deeper into the star. A similar argument could be made to show why water pressure must increase at greater depths in the ocean; the higher pressure supports the overlying layers.

In a careful treatment of stellar structure we would consider each infinitesimal shell of gas, calculating the pressure needed to provide support down through the star. Such calculations show that the larger the total mass of the star, the greater the central pressure. We are also interested in the temperature structure of a star, since this determines nuclear reaction rates, the interactions of photons with the star's ionized gas, and so forth. For normal stars, the ideal gas law provides the relationship between pressure and temperature that we need. We might be concerned that this simple law would fail for the extreme conditions in the deep interior of a star, but real gases actually obey the law to an excellent degree even at very high temperatures. Applying the ideal gas law tells us that the more massive the star, and hence the higher the central pressure, the hotter it must be at its core.

Although it plays by far the major role, the mass is not the only quantity important to a star. It is slightly less straightforward to visualize, but the elemental composition also affects a star's structure, through the average mass per gas particle. When the average particle mass is higher, then fewer particles are present for a given total stellar mass. This means the number density n is lower, which in turn implies that the temperature must be higher to produce a given central pressure. Most stars have similar compositions, consisting of approximately

three-quarters hydrogen and one-quarter helium, by mass, with other elements present in small quantities. However, the compositional variations from star to star, even though relatively small, can produce subtle differences. A careful analysis of the equations of stellar structure shows that the nature of a star is almost entirely controlled by its mass and its composition.

Up to this point we have concerned ourselves with the implications of hydrostatic equilibrium. If this were the whole story, the star would exist indefinitely in a static state. But stars radiate heat and light into space; were it not for nuclear reactions that replenish the lost energy, the stars would cool and go out of equilibrium. The high temperature in the stellar core is just what is needed to drive those reactions.

Twinkle, twinkle, little star

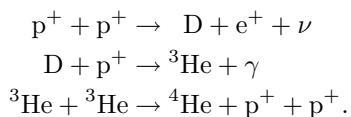
Why is the temperature so important in nuclear fusion? The core of a star is composed predominantly of free protons and electrons whipping around at very high speeds. Under conditions even close to what we Earthlings might regard as ordinary, two protons repel one another, since both have positive electrical charge. The closer the protons approach, the more strongly they repel one another, because the electrostatic force follows an inverse square law. This mutual repulsion creates the *Coulomb barrier*, which ordinarily keeps the protons apart. In an atomic nucleus, however, protons manage to stick together despite the electrostatic repulsion. This is possible because the protons are bound together by another force, the strong nuclear force, and this force is much stronger than the electrostatic force. But the nuclear force has a very short range, comparable to the diameter of the nucleus. The trick, then, is to force the protons sufficiently close together that the nuclear interaction can take over. The higher the temperature, the closer the protons can approach. In the core of the star, the protons have very high energies (temperature), and are forced extremely close together. Under such conditions, occasionally a purely quantum effect called *tunneling* can occur; the protons pass through the Coulomb barrier and merge. The product of this fusion is not two protons stuck together, but is a *deuteron*, the nucleus of the **deuterium**, or heavy hydrogen, atom; the deuteron consists of one proton and one neutron. When the protons fuse one proton is converted to a neutron, and a positron and a neutrino are ejected from the new nucleus.² The positron immediately annihilates with an electron, releasing energy; the neutrino also carries away some energy. Neutrinos interact so little with ordinary matter that the energy they carry is essentially lost immediately from the star. Overall, an amount of energy equal to the *binding energy* of the deuterium nucleus

Nuclear fusion

²Technically, this reaction also involves the *weak interaction*; the weak interaction is distinct from the strong interaction but also operates only at the scale of the atomic nucleus. For the present purposes, however, the general picture we have developed here is adequate.

is liberated. The binding energy is the amount of energy required to break apart the nucleus; hence when such a nucleus forms, the same amount of energy must be released.

Deuterium is only too happy to fuse with another proton to form ${}^3\text{He}$, releasing a high-energy photon (γ). The new ${}^3\text{He}$ nucleus quickly reacts with another to create ${}^4\text{He}$, a very stable nucleus; two protons are also produced, which may then re-enter the cycle. The net result is the fusion of four protons into one nucleus of ${}^4\text{He}$, the creation of two neutrinos, and the liberation of the binding energy of the helium nucleus. The total mass-energy released by fusing hydrogen to helium is about 0.7% of the rest mass-energy of the reactants. Schematically, we can write the reactions involved in the *proton-proton process* as



There are additional reaction routes that convert the ${}^3\text{He}$ into ${}^4\text{He}$, but the final result is largely the same.

To give a specific example, let us calculate how much hydrogen burning is required to account for the energy emitted by the Sun. The Sun has a luminosity of $3.9 \times 10^{26} \text{ J s}^{-1}$. Dividing this luminosity by the fraction of its rest mass that is converted to energy, $0.007c^2$, yields $6 \times 10^{11} \text{ kg}$ of hydrogen per second, or about 600 million metric tonnes. The energy thereby released slowly makes its way in the form of photons to the outer layers of the star. The way is difficult, for the hot inner layers are opaque, and photons are constantly scattered, absorbed, and re-emitted. In the Sun, a star of average density, a photon generated in the core takes hundreds of thousands of years to work its way to the transparent outer layers; from there it can at last stream into space. The light falling on us today was generated in nuclear reactions in the Sun's core that occurred before our kind walked the Earth.

The CNO cycle

There is another process, the *CNO cycle*, in which carbon-12 goes through reactions with protons, passing through nitrogen (13, 14, and 15) and oxygen-15 before the last step, in which nitrogen-15 fuses with a proton and emits an alpha particle, that is, a nucleus of ${}^4\text{He}$, thereby reverting to carbon-12. Although it is much more complicated than the proton–proton process, the net result of the CNO cycle is the combination of four protons to create one nucleus of ${}^4\text{He}$, along with the emission of two positrons and two neutrinos. The carbon-12 re-emerges at the end unchanged; it thus functions as a *catalyst*, a substance that participates in and assists a reaction, but itself is unaffected overall. The rate at which the CNO cycle proceeds is highly temperature sensitive, and it is rare in stars like the Sun; it is important only for stars more massive, and thus hotter, than the Sun. The CNO cycle also, obviously, depends upon the presence of carbon atoms. We shall eventually learn that only hydrogen and helium, and a small fraction of the light element lithium, are **primordial elements**, those elements that were created near the

beginning of the universe, before the first stars formed. With a very few exceptions all other elements are manufactured in stars. The CNO cycle thus depends upon the existence of earlier generations of stars that made carbon; it was not available to the first stellar generation. Most elements heavier than helium are manufactured by stars over the course of their lifetimes, as they fuse one nuclear fuel after another in an attempt to maintain their structures against the pull of gravity. The newly created elements within a star return to interstellar space when the star sheds most of its gas at the end of its life. There the enriched gas may join other clouds of gas to bring forth later-generation stars, such as our own Sun. The oxygen we breathe, the carbon and nitrogen and sulphur and phosphorus that make up much of our bodies, the iron and aluminum and the silicon upon which our industries and economies are based, indeed, almost all of the matter on Earth, and in our own bodies, was created within ancient, massive stars that lived and died before the Sun was born.

The details of the nuclear processes are not as important as is the realization that they provide the energy to keep the star in hydrostatic equilibrium. We can make some further progress in understanding stars without any knowledge of nuclear reaction rates. The mass of a star is the most important factor in establishing its core temperature. Temperature, in turn, determines the rate at which nuclear reactions proceed in the star's core. The energy released in the core must work its way through the star to be released at its surface, thus ultimately determining the star's luminosity. It follows that there must be a relationship between the mass of a star and its luminosity. The ingredient needed to complete that relationship is an approximate relationship between temperature, luminosity, and stellar radius. This in turn depends upon the rate at which energy can be transported through the star. A very simple physical argument, which assumes that photons diffuse through the dense gas deep within the star till they reach the thin outer layers, yields an approximate relationship between luminosity and mass of

$$L \propto M^3. \quad (5.3)$$

For example, a star of ten solar masses, $10 M_{\odot}$, would have a luminosity not 10 times, but 1000 times that of the Sun. The luminosity, and hence the energy consumption, of a star thus goes up quite rapidly with its mass.

So far, we have discussed only theory; what about observations? There is a straightforward relationship between the luminosity of a star and its *surface* temperature, T_s . Stars are nearly blackbodies, so the energy per unit area they radiate is proportional to the fourth power of the surface temperature, T_s^4 . The total luminosity will be the energy per unit area times the total surface area, which is proportional to R^2 ; hence $L \propto R^2 T_s^4$. Luminosity cannot be observed directly, but surface temperature is relatively easy to measure. It is only necessary to observe the continuum spectrum of the star, determining where that spectrum peaks; the blackbody relationship then gives the surface temperature.

The mass of a star determines its luminosity

Stellar temperatures can be measured by observing a star's color

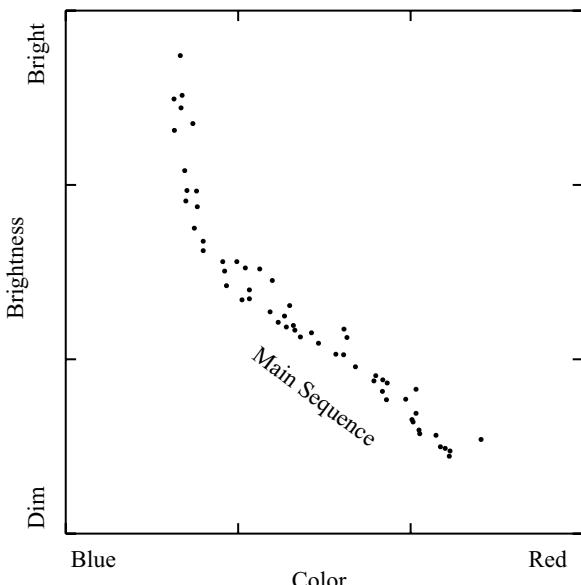


Fig. 5.6 Composite Hertzsprung–Russell diagram for some of the stars of the Pleiades, a young stellar cluster. The vertical axis is the logarithm of the brightness of the star; for stars at a fixed distance the observed brightness will be proportional to the luminosity. The color is a measure of the temperature of the star. The points do not appear randomly, but lie along a curve called the main sequence. Stars on the main sequence fuse hydrogen to helium in their cores.

(If that is not sufficiently accurate, known corrections can be applied to make a better model of the radiation of the star, and an improved value for the temperature can then be computed.) The color of a star is related to its surface temperature. The redder the star, the cooler its surface. Bluer stars are hotter. If the distance to the star can be measured by independent means, then the observed flux can be converted into total luminosity. Another approach is to study a group of stars at the same distance, such as a star cluster. In either case, it is possible to measure the luminosity and the temperature for a number of stars and make a plot. The plot should reflect the underlying stellar physics we have described with our simple physical stellar models.

If the surface temperature of the star is plotted as a function of its luminosity, we obtain a graph called the *Hertzsprung–Russell diagram* (generally shortened to the HR diagram). The points are not scattered about, but fall into very narrow and well-defined curves. Most stars lie on the **main sequence**. Along the middle portion of the main sequence, the luminosity is related to the mass by $L \propto M^{3.1}$, very close to the value obtained by a simple physical argument. Thus, the observed main sequence seen in the HR diagram indicates that the processes occurring in stars are controlled primarily by the conditions required for hydrostatic equilibrium, the balance between gravity and the pressure supplied by the heat from nuclear reactions. For as long as the fusion of hydrogen to helium dominates, the star resides, usually quietly, on the main sequence of the Hertzsprung–Russell diagram.

If we wished to develop a realistic stellar model, we would have to write down the differential equation of hydrostatic equilibrium. Then we would be obliged to include rate equations for the nuclear reactions in the core, and we would be required to solve the difficult equations of

The HR diagram reveals relationships between stellar temperature and luminosity

radiative transfer. A realistic model of a star is sufficiently complicated that it is necessary to solve the resulting equations with a computer using numerical techniques. Although not all of the phenomena are perfectly understood, especially those having to do with the transport of energy within the star, stellar models are still good enough to reproduce the main sequence to a high degree of accuracy; stars are probably among the best-understood structures in the universe. They are very important to cosmology, because their lives are uniform and predictable. And their ages, and their deaths, have significant ramifications for the universe and its contents.

Stellar ages

Astrophysicists like to joke that “we understand every star except the Sun.” The problem with the Sun, of course, is that we have an overwhelming amount of data on every detail of its existence, including its every magnetic outburst and minor shudder. We cannot forecast, or sometimes even explain, the day-to-day workings of the Sun. Even so, we do understand the fundamentals of the construction of the stars, and the grand outlines of their lives. We can exploit this knowledge to determine the ages of stars, by comparing observations to the predictions of our models.

The stringent physical constraints that govern the evolution of a star result in a predictable life history. Stars wage a constant battle against their tendency to collapse under their own weight. The nuclear reactions deep within the interior provide the energy that is radiated away by the star; as long as the lost energy is constantly replenished by fusion, the temperature at the core can be maintained high enough to fight the inexorable pull of gravity. The great majority of a star’s existence is spent on the main sequence as a hydrogen-burning star; for all practical purposes, we can define the lifetime of a star to be its time on the main sequence. Recall that main sequence stars have a luminosity-to-mass dependency of roughly $L \sim M^3$. The life expectancy of a star is determined by the amount of fuel available to it, divided by the rate at which it consumes that fuel. Hence the stellar lifetime is proportional to its mass (fuel) divided by its luminosity (burn rate), that is, $t_* \sim M/L$. Together, these relationships imply that stellar lifetimes decrease with the square of increasing mass, $t_* \sim M^{-2}$. This is just a rough calculation, but it indicates that massive stars live much shorter lives in comparison to low-mass stars. Very massive stars burn their candles at all ends, blazing gloriously for a few million years before exhausting their supplies. Less massive stars, such as the Sun, burn their fuel more frugally and exist in a stable state for many billions of years. This fact can be used to set a limit on an important cosmological measurement: the age of the universe.

Stellar lifetimes are determined by mass

Studies of stellar ages have determined that the stars of the Milky Way Galaxy fall into two broad categories, called **Population I** and

Population II. Population II (or just Pop II) stars are very old, probably nearly as old as the Galaxy itself, whereas Population I (Pop I) stars like the Sun are much younger, and continue to form today. The major difference between the stellar populations, other than their ages, is their composition. Old stars have far fewer **metals**, which to astronomers means any element heavier than helium, whether chemically a metal or not. This is consistent with the formation of heavy elements within stars; the early generations of stars must have formed from gas that had little metal content, since there were few earlier stars to create the metals. Population I stars condensed from the debris of older, massive stars that exhausted themselves quickly, and they and presumably their environs are considerably enhanced with metals.

There is speculation that there exists a primordial population of ancient pre-galactic stars, called **Population III**. These stars would be the very first formed in the cosmos. There is currently no definitive evidence for their continued presence today, and there are theoretical arguments that most such objects would have been sufficiently massive that they would have long since exhausted their fuel and died. However, one object that might belong to this mysterious population was found in 2002 in the Galactic halo. Its metallicity was a minuscule 1/200,000th that of the Sun, but, even so, its surface was slightly enriched with carbon and nitrogen that probably came from a long-ago companion. However, even with this discovery, Population III stars are still largely hypothetical.

While most Population I stars are found in the spiral disk of the Milky Way, most Population II stars are found in the Galactic bulge toward the center of the Galaxy, or in the halo surrounding the disk of the Galaxy. In the halo they are often found in **globular clusters**, huge, nearly spherical clusters of about 100,000 stars each. The globular clusters orbit the Galactic center within a roughly spherical volume. They are found not only around the Milky Way, but are also seen around every other galaxy close enough for objects of their size to be resolved. Globular clusters are thought to be the oldest objects in the Galaxy; thus the age of their most ancient stars provides a lower limit to the age of the Galaxy, and hence of the universe itself.

Consider such a cluster of stars, whether a globular cluster or a younger open cluster. As discussed previously, the cluster stars formed at about the same time, from the same nebula. Thus the stars should all have about the same initial composition. Stars of all masses were created, in accordance with the initial mass function. An HR diagram of the cluster would reveal a full main sequence distribution of these stars. As time goes by, the most massive stars use up their hydrogen and evolve off the main sequence. Slowly, the main sequence disappears, starting at the high-mass end and moving toward the low. Of course, we can observe only a snapshot of a star cluster at one particular time in its evolution; if we plot an HR diagram of all the stars that are members of the cluster, we will find many dots spread along the main sequence, with an abrupt cutoff corresponding to those stars that are on the verge of ending their main sequence lives. (Stars more massive than this have

Primordial stars

Determining a star cluster's age

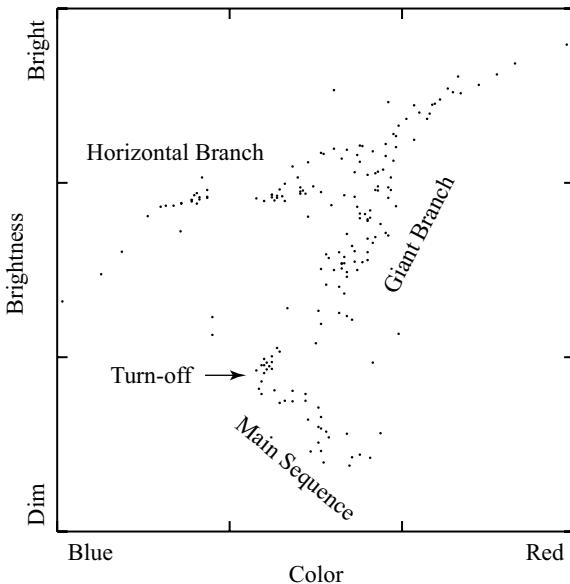


Fig. 5.7 Composite Hertzsprung–Russell diagram for some of the stars of M3, an old globular cluster, plotted in units of brightness versus color. Only the stars in the lower center of the plot lie along the main sequence. Stars that have left the main sequence lie above, on the horizontal branch, and to the right, on the giant branch. (Compare with Figure 5.6.) The point at which the stars leave the main sequence is the turnoff point. A comparison of the observed main-sequence turnoff with the predictions of stellar theory gives the age of the cluster.

already left the main sequence.) The cutoff point represents a specific mass, the **turnoff mass**. If the main sequence lifetime of stars of that mass is known, then the age of the cluster is determined.

There are, of course, many uncertainties in age determinations. Variations in composition, mass loss, and the effects of turbulent mixing in stellar layers are examples of potential sources of error. Much of the uncertainty lies with unknown stellar compositions; there are also difficulties in determining precise main sequence turnoff points, and matching those points with theoretical models. Even so, experts in stellar ages have reached a consensus. The oldest globular clusters in the Milky Way Galaxy and its neighbors have been determined to be from 12 to 15 billion years old. At the present time, it appears to be quite difficult to find a reasonable combination of error and uncertainty that would produce stellar ages in the oldest globular clusters of less than about 12 billion years.

White dwarfs to black holes

For stars in the early and middle stages of their lives, the most important nuclear reaction is the fusion of hydrogen into helium. But all such things must end, and every star eventually runs out of usable hydrogen fuel. For example, if the Sun began its life entirely composed of hydrogen and gradually converted all this hydrogen to helium, it could live for 100 billion years. However, nuclear reactions take place only where the temperature and density are high enough, and in the Sun this is the case only deep within its core. At best, the Sun can convert no more than about 10% of its hydrogen to helium. When hydrogen can no longer be

The end of a star's life

*The red giant phase**Formation of a white dwarf*

fused, the pressure in the core drops, allowing gravity again to compress the gas. The core contracts and the star changes its structure. As it contracts, the temperature in the core rises. This increase in the core temperature is important, because the next nuclear fuel to be burned, helium, does not fuse at the lower temperatures found during the star's main sequence life. When the core becomes sufficiently hot, helium will begin to fuse into carbon in the deep interior, while hydrogen continues to burn in a relatively thin shell surrounding the core. The new fusion reaction is called the *triple-alpha* process, because three nuclei of ^4He fuse to ^{12}C , with the unstable nucleus ^8Be (beryllium-8) created as an intermediate product. This new energy source stops the gravitational contraction and stabilizes the star, allowing it to continue to shine, once again in equilibrium. The star exits the main sequence for the *horizontal branch*, the region above and to the right of the main sequence on the Hertzsprung–Russell diagram that is occupied by stable helium-burning stars. The increase in temperature causes the outer regions of the star to expand, increasing its luminosity, and the expansion cools the surface layers, which shifts the radiation to lower energies. The star balloons to enormous size, creating a **red giant**. When the Sun reaches the red giant phase, perhaps five billion years into the future, its surface will extend to near the orbit of the Earth. The intense radiation falling upon the Earth will destroy any life that might remain, and the planet itself will spiral into the bloated Sun, vaporizing in its hotter inner layers. The Sun will reclaim its innermost planets.

But this stage can only last so long, for the helium is consumed even more quickly than the hydrogen before it. What happens next depends upon the mass of the star. For stars of modest mass, such as the Sun, the end is quiet. The heavier the nucleus, in general, the higher the temperature and density required to force it to participate in fusion reactions. Stars up to about 6 times the mass of the Sun are not sufficiently massive for gravity to be able to raise the core temperature to a high enough level for further fusion reactions to occur. Once the usable helium fuel has been converted to carbon in such stars, nuclear reactions cease and the core once again contracts under its own gravity. The contraction continues until the matter in the core becomes so compact that electrons cannot be squeezed together further. This state is called **electron degeneracy**, and it is a quantum mechanical consequence of the Pauli exclusion principle. Electrons are fermions and thus, by the exclusion principle, no two can occupy the same quantum state. In electron degeneracy, all low-energy quantum states are occupied, forcing many electrons into high-energy states. It would take considerable energy to squeeze the electrons even closer together, so the electrons provide a new source of pressure that does not depend on temperature. This is quite different from the ideal-gas law; it is somewhat analogous to the intermolecular electrostatic forces that give a crystal such as quartz or diamond its great rigidity. Most significantly, it means that the star can now resist gravity with no further generation of heat. As the core settles down to its degenerate state, nuclear burning can continue in the

surrounding stellar envelope. This eventually causes the star to eject its swollen outer layers; if we happen to observe the expanding shell of gas, it might take the form of a lovely object called, for historical reasons, a *planetary nebula*. Eventually, only the degenerate core is left behind as a stellar ember known as a **white dwarf**. A white dwarf star no longer burns nuclear fuel, and shines only because it takes many millions of years for light to percolate out to the surface from deep within its core. Eventually the star will cool, and the white dwarf fades away. This is the eventual fate of our Sun.

White dwarfs have sufficiently low luminosity that the only ones we can observe directly are in our solar neighborhood. The bright star Sirius is actually a binary; the tiny companion, invisible without a good telescope, is a white dwarf, the first discovered. All white dwarfs are very small and very dense. (A white dwarf with the mass of the Sun would be packed into a volume the size of the Earth.) This immediately tells us that their gravitational fields are relatively strong. The chemical composition of white dwarfs probably varies somewhat, but observations are consistent with the theory that they should consist predominantly of carbon, with some oxygen, the final products of helium burning. The unusual state of the matter in a white dwarf has some interesting consequences. For one thing, the greater the mass of a degenerate white dwarf, the smaller its radius. For another, as the white dwarf cools, it can actually *crystallize*; its nuclei, long separated from their electrons, behave much more like a solid than like the gaseous plasma of which the star was previously composed.

Since a white dwarf is no longer generating energy, it cools at a rate determined mainly by only a few quantities: its surface temperature and area, which control the rate at which energy is radiated into space, and the length of time required for a photon to work its way from the interior to the surface. White dwarfs have extremely high surface temperatures, as much as tens of thousands of degrees, but not a lot of surface area, so overall they radiate rather slowly. Moreover, they are so dense that it takes a very long time for photons to diffuse outward. As photons slowly trickle to the surface of the white dwarf and stream away, the star loses energy and cools; with time, a white dwarf will shift its color from blue-white to yellow to red, and then will finally cease to emit in the visible at all. White dwarfs cool so slowly, however, that the universe is probably still too young for a significant number to have disappeared from visibility.

White dwarf cooling

If we could compute the rate of cooling of a white dwarf, we could deduce the time elapsed since it formed. In principle, this is quite possible; in practice, there are many difficulties. Since we cannot fetch any samples of white-dwarf matter, nor can we recreate it in the laboratory, we must rely upon theory to construct models of the characteristics of the material of which the dwarfs are composed, then attempt to compare the predictions of the model with observations of real white dwarfs. Another unfortunate limitation is that our sample of white dwarfs is small. They are dim and tiny—the brightest have a luminosity of approximately

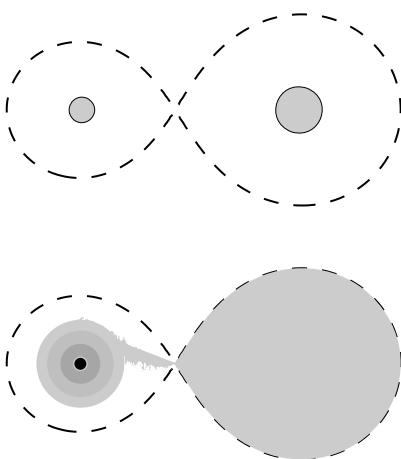
Interacting binary systems

Fig. 5.8 The Roche lobe (dashed line) surrounding binary stars is the surface that marks the region within which a particle is bound to one star or the other. The crossing point in the figure-eight is the point at which the two stars' gravitational attractions are equal but oppositely directed. If one star fills up its portion of the Roche lobe, gas can overflow and transfer to the other star, creating an accretion disk around the companion.

$0.1 L_\odot$ —and we can see only those in our Galactic neighborhood, even with the best of modern telescopes. Most of those we can find are the binary companions of normal stars. Nevertheless, many efforts have been made to estimate the age of the oldest white dwarfs in the Galaxy, since such a datum would obviously set a lower limit to the age of the Galaxy itself. The best estimates obtain an age of approximately 12 billion years for the most ancient white dwarfs, consistent with the ages of the globular clusters.

Occasionally, a white dwarf can revive if it has a companion. When two stars orbit, their gravitational fields overlap, since gravity's range is infinite. Each of the stars is surrounded by a region, called the *Roche lobe*, within which its gravity dominates that of its partner. The Roche lobes of the members of a binary touch at a point known as the *Lagrange point*; this is where the gravitational tug of each star is equal in magnitude. In a typical binary system, each star is much smaller than its Roche lobe. If the separation between the two is large, both stars will spend their lives well within the confines of their Roche lobes. However, as stars age and leave the main sequence, they swell to giant size. In a close binary system, when one of the stars reaches the red giant phase it can overflow its Roche lobe, and some of the distended star's outer layers can be transferred onto the smaller companion. Since the members of a binary star system are in mutual orbit around one another, any gas flowing from one star must partake of this rotation. Thus we encounter a situation reminiscent of the formation of a protostellar disk; as gas flows from one star to the other, it falls inward along an orbital trajectory. If the star toward which the gas is falling is small enough, and this will certainly be true for a white dwarf, the inflowing gas stream misses the star's surface and goes into orbit. In this case, the inflowing gas creates an **accretion disk**. Dissipation in the gas through turbulence in the disk means that a parcel of gas cannot orbit its new primary at a fixed radius. Rather, it spirals toward the star. The fall of the gas in the gravitational field releases energy; the gas in the disk is compressed and heated. The accretion disk may emit high-energy radiation, even X-rays, which can be detected from the Earth. Eventually, the gas crashes onto the surface of the star, emitting a burst of energetic radiation.

If the accreting star happens to be a white dwarf, the transfer of mass can have some other interesting consequences. A white dwarf cannot incorporate the new material in a smooth manner, as would a normal star, since a dwarf's pressure support comes not from ordinary gas pressure but from degeneracy pressure. Whereas ordinary pressure can adjust with changes in temperature and density, degeneracy pressure is independent of temperature, and no adjustment occurs when new gas impinges upon the white dwarf. The infalling gas is thus compressed and heated as it strikes the unyielding surface of the white dwarf. When enough gas has piled up, it can reach the 10^7 K required to trigger hydrogen fusion. The white dwarf suddenly flares in brightness and becomes a **nova**. After this thermonuclear explosion from its surface, it once again fades away. Often, the cycle is repeated, when enough gas

again accumulates to reach the ignition point. At their peak brightness, novae seem to have fairly uniform luminosities, which means it might be possible to use them to determine distances. Unfortunately they are not perfectly standard; work continues to determine whether novae can help to calibrate the cosmic distance scale.

What is the fate of stars that cannot settle down to a quiet retirement as a white dwarf? Stars more massive than a few solar masses experience more phases at the ends of their lives, going through one nuclear fuel after another to battle the crush of gravity. After the star's helium is exhausted, the core contracts and heats again and the outer layers expand, sending the star up the *asymptotic giant branch* of the Hertzsprung–Russell diagram. In very massive stars, carbon may first ignite; for sufficiently massive stars, increasingly heavy elements are subsequently burned, fusing all the way to iron. The star becomes a gigantic cosmic onion, consisting of concentric shells in which increasingly heavier elements are fused. The final fusion product is iron. Unlike the lighter elements, iron demands an *input* of energy to be forced to fuse into heavier elements. Iron cannot provide the energy the star needs in order to support its weight, since further fusion would actually consume the star's precious energy. Once the available matter has been fused to iron, the star is out of usable fuel; iron is the end of the nuclear road.

A star with an iron core must seek an equilibrium with gravity that does not require further expenditure of energy. Smaller stars could find their final state in electron degeneracy. However, Subramanyan Chandrasekhar realized in 1930 that in order to provide the incredible pressures required to maintain stars more massive than $1.4 M_{\odot}$, the degenerate electrons would have to move at greater than the speed of light. This was known from the special theory of relativity to be an impossibility; thus, special relativity demands an upper limit to the mass of a white dwarf. This bound is called the **Chandrasekhar limit**. For masses above this limit, the pressure from even electron degeneracy is not sufficient to support the star against its own weight. If the dying star fails to eject enough of its matter to allow its collapsing core to drop below this limit, the electrons cannot supply the necessary pressure. But if electron degeneracy pressure falls short, the star does not just slowly contract. It collapses catastrophically, sending a shock wave into its outer layers and blowing them off in a single cataclysmic explosion called a **supernova**. A supernova is so bright that for a brief interval of a few weeks, it may outshine the galaxy in which it occurred, a blazing beacon visible across enormous distances. The explosion is so powerful that most of the star's matter is blasted away and dispersed into space. This may seem to be a cruel finale for a star, but the supernova plays a vital role in the history of the cosmos. It is in supernovae that the heavier elements, forged near the center of the star during its last stages of existence, can find their way into space, and thence into later stars and planets. Indeed, so much energy is liberated in the blast that elements heavier than iron can be created, even though, as remarked above, these reactions consume energy rather than releasing it. The gold and silver with which we

Very massive stars fuse increasingly heavy elements in their cores as they age

The Chandrasekhar limit

The death of a massive star is a spectacular supernova

ornament ourselves, and which we hoard and covet, came into abrupt being in the final moments of the life of a massive star. Cobalt and nickel and zinc, and the uranium of our nuclear-power plants and of our weapons—all these heavy elements are the ashes of massive stars.

Type I and Type II supernovae

A supernova that arises from the collapse of a massive star is designated by astronomers as Type II. This suggests that there must be another type of supernova, the Type I supernova. The Type Ia supernova represents the explosion of a white dwarf in a binary system.³ As we have seen, the accretion of matter from a binary companion onto a white dwarf can lead to a nova explosion. A nova outburst probably blasts away much of the gas accumulating on the surface of the dwarf, but not all of it. After each cycle of nova activity, the mass of the white dwarf may increase slightly; eventually, it may acquire more gas than it can support. It is constrained by the Chandrasekhar limit throughout its existence. Should its mass rise above that limit for any reason, including mass transfer from a binary companion, electron degeneracy pressure cannot continue to support it. The star collapses violently. The sudden increase in temperature detonates the carbon; because of the degeneracy of the matter, *all* the matter in the white dwarf fuses almost simultaneously. The resulting explosion rivals the death of the supermassive star in its brilliance.

The importance of Type Ia supernovae for cosmology

Type Ia supernovae have a property that is of particular interest to the cosmologist. In a Type Ia supernova, the progenitor was very near to the Chandrasekhar limit, else it would not have collapsed in the first place. Therefore, Type Ia supernovae tend to peak in energy output at very similar luminosities. In fact, there exists a relationship between the peak luminosity and the rate at which the light from the explosion fades away. This means that by studying the light from a Type Ia supernova as it brightens and fades, astronomers can determine their intrinsic luminosity and use them as distance indicators. Since we need reliable standards by which to calibrate the distance scale, a supernova's utility is greatly enhanced by its brilliance. They can be seen easily in very remote galaxies. Observations of such distant supernovae have had remarkable implications for our understanding of the universe.

The formation of a neutron star

Although a great deal of the star is blown out into interstellar space by a Type II supernova, some fraction is probably left behind in a core remnant. If the mass of the remnant still exceeds the Chandrasekhar limit, what can the star do? It cannot settle down as a white dwarf star; so what remains? As the star collapses to greater and greater compaction, the electrons are squeezed into the atomic nuclei themselves, where they are forced to merge into the protons, forming neutrons. The neutrons, which are much more massive than electrons, can themselves exert a degeneracy pressure known as **neutron degeneracy** pressure. The entire star is compressed essentially to the density of an atomic nu-

³What distinguishes a Type I from a Type II observationally is the absence of hydrogen lines in the supernova's spectrum. There are other, rarer, kinds of Type I supernovae that are not due to white dwarf explosions. For the present purposes we can ignore those other varieties.

cleus, but composed only of neutrons. This massive neutron nucleus is known as a **neutron star**. These objects are astonishingly compact; a neutron star with the mass of the Sun would have a radius of only about 10 kilometers, roughly the size of a typical large city on Earth. The neutron star is a remarkable object. Its existence was predicted as early as 1934 by Fritz Zwicky and Walter Baade, who proposed that neutron stars could be created in supernovae. The first detailed calculations of neutron star structure were performed by J. Robert Oppenheimer and George Volkoff in 1939. The work of Baade, Zwicky, Oppenheimer, and Volkoff was largely ignored for decades; such a star seemed too bizarre to consider. This attitude changed in 1967 when the first **pulsar** was detected. A pulsar emits highly regular, energetic bursts of electromagnetic radiation, generally as radio waves. The pulses from the first pulsar were so regular that the discoverers, Jocelyn Bell and Anthony Hewish, first thought they had received signals from another civilization! No familiar astronomical process was known at the time that could produce electromagnetic bursts of such sharpness and regularity, at such a rapid rate. Ordinary oscillations would be inadequate to explain the signal. As more and more pulsars were observed, however, the mystery slowly yielded. Thomas Gold first suggested that pulsars might be associated with the exotic neutron star. Subsequent observations have borne this idea out; no other mechanism is remotely plausible to explain the properties of pulsars.

Astronomers have learned much about neutron stars, but their structure is still somewhat mysterious. The matter in a neutron star is compressed into an even stranger state than that of a white dwarf. A white dwarf is somewhat like a very dense solid; unusual, but not mind-boggling. A neutron star is much weirder, more like a huge atomic nucleus than it is like anything familiar. The interior of the neutron star is probably in a fluid state, meaning that the neutrons move around freely. They move so freely, in fact, that the interior is said to be a *superfluid*, a fluid in which no friction is present. The fluid of degenerate neutrons is surrounded by a thin crust of fairly normal matter, consisting of crystalline iron nuclei, free electrons, and free neutrons. The tiny radius and the large mass of a neutron star imply an enormous, almost incomprehensible, gravitational field near its surface. Occasionally, the intense gravitational field causes a defect in the crystalline structure of the crust to crack, resulting in a *starquake* as the crust readjusts. The starquake causes a *glitch* in the pulsar, a small but very sudden drop in the period of its pulsation. These starquakes provide valuable information into the nature of neutron star matter.

A neutron star is no longer generating energy from fusion; how might it send pulses of energy into space? Suppose a hot spot is present on the surface of a rotating star. The light emissions from such a spot would sweep through space like the beacon from a rotating lighthouse lamp. Just as a sailor sees the beam from the lighthouse only when it points at him, so we see the radiation from the pulsar at intervals equal to the rotation period of the star. But what kind of star could

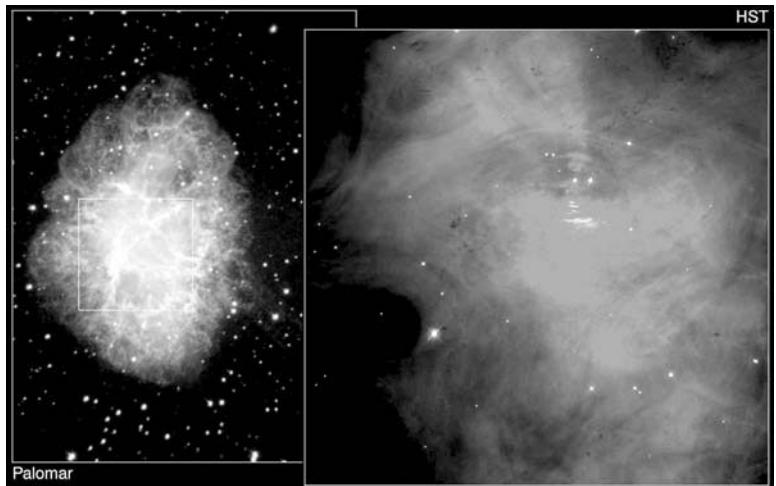


Fig. 5.9 Images of the Crab Nebula. On the left is a ground-based wide-field view of the entire nebula, showing the shocked filaments. On the right is a photograph of the innermost region by the *Hubble Space Telescope*. The central pulsar, one of the few visible in optical wavelengths, is visible at the center of the nebula. (The bright star just to its right is not associated with the nebula.) Due to its proximity and young age, the Crab Nebula has provided much important information about supernovae and their remnants. (J. Hester and P. Scowen, Arizona State University; NASA/STScI.)

rotate once per second? If it were a white dwarf, about the size of the Earth, such rapid rotation would tear it apart. A neutron star, on the other hand, would have only about the diameter of a typical city, and could easily rotate at such a speed without breaking up. The case was clinched in 1968 by the discovery of a pulsar in the center of the famous Crab Nebula, an untidy blob of gas in the constellation Taurus. The Crab Nebula is well identified with a supernova observed in AD 1054 by Chinese astronomers; it is the shocked, disordered remnant of the outer layers of the star. Thus, the association between a pulsar in the Crab Nebula and the known supernova that had occurred there made the identification of pulsars with neutron stars all the more certain. The Crab pulsar emits approximately 30 pulses per second, one of the most rapid rates of any pulsar. It pulses in optical wavelengths as well as radio, also unusual. Since the date of the supernova is known, we can conclude that the Crab pulsar is young. Fast pulse rates and high energy output are associated with recently formed pulsars. As they age, they lose rotational energy and the period of their pulsations increases.

If a neutron star is such a dense, exotic object, how could it be set into such rapid spinning? It is a consequence of an important law of physics, the **conservation of angular momentum**. Angular momentum is a measure of the resistance of a body to changes in its rotation, and is defined as

$$\mathcal{L} = I\omega, \quad (5.4)$$

where ω is the rotation rate of the body, in angle per unit time, while I is a quantity called the *moment of inertia*. The moment of inertia describes the matter distribution of the object; the farther the mass from the axis of rotation, the greater the moment of inertia. Conversely, the more concentrated the matter near the axis of rotation, the smaller the moment of inertia. The law of conservation of angular momentum states that if no outside torque, or twisting, acts upon the body, its angular momentum does not change. Therefore, if the moment of inertia

Conservation of angular momentum

changes, the rate of rotation must change in such a way that the angular momentum remains the same. Perhaps the most familiar illustration of the conservation of angular momentum is the figure skater executing a spin. The skater usually begins with arms outstretched, spinning at a certain rate. As he draws his arms toward his body, his moment of inertia decreases; to conserve the total angular momentum, there must be an increase in his rate of spin. Occasionally the skater even crouches, pulling all parts of his body close to the axis of rotation to increase his rate of spin even further. As he unfolds, his moment of inertia increases and his spin decreases, until he is spinning slowly enough to stop easily by exerting a small torque with the skate blade. Another everyday example is the ability of cats to land on their feet most of the time. Even a falling cat must obey the law of angular momentum as it rights itself. High-speed photography of cats dropped from safe distances clearly shows them to twist their front legs in one direction, while their hind legs twist oppositely. The cat is still able to turn its body to land feet downward, but at each motion its angular momentum must be conserved as it falls.

It is no accident that the moment of inertia is reminiscent of the inertial mass; its role for rotational motions is analogous to the function of inertial mass for linear motions. Since the moment of inertia of a star depends upon its mass distribution, the gravitational collapse will change the moment of inertia drastically. The radius of the core of the dying star can shrink abruptly by a factor of perhaps a thousand or more. For a sphere, the moment of inertia is given by the formula

$$I_s = \frac{2}{5}MR^2. \quad (5.5)$$

Thus, if little mass is lost, the newly formed neutron star must spin approximately a million times faster than its precursor. Typical pulsars rotate with periods of one second to approximately a quarter second. For the idealized example of a neutron star executing one rotation per second, conservation of angular momentum would imply that the precursor rotated about once per month, which is comparable to the rotation rate of the Sun.

Along with rotation, the neutron star must possess a hot spot in order to emit the beamed radiation. A lighthouse mirror may turn, but unless the lamp is lit, there will be no beam. How does the hot spot generate such radiation? As far as we know, all stars possess magnetic fields; the field is tightly coupled with the ionized gas of which the star consists. When the core of a massive star collapses, its magnetic field is pulled along with it, greatly concentrating the field and producing huge magnetic forces. Most astronomical objects, including the Sun, have overall magnetic fields that look somewhat similar to, though they are stronger than, the field of a bar magnet; there is a north and a south pole, and the field lines run continuously from one pole to the other. The collapse probably does not change the basic configuration of this field, although it does greatly amplify it, since the magnetic field is tied to the matter and becomes stronger as the radius decreases and the density

Pulsar rotation

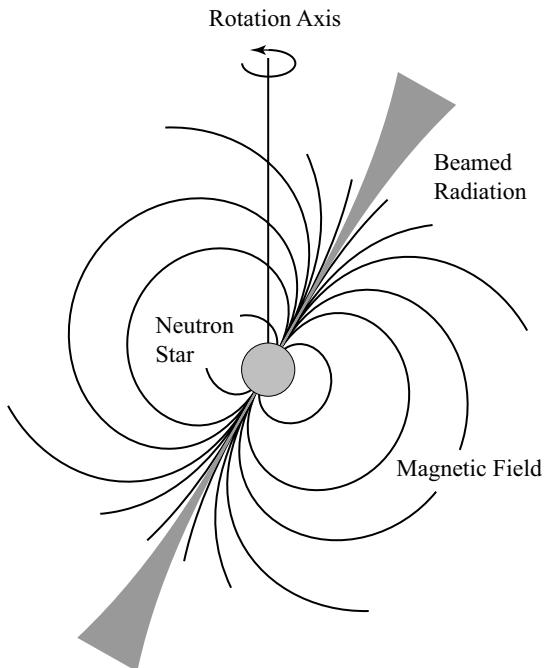


Fig. 5.10 A pulsar has a hot spot that is carried around by the pulsar's spin. The location of the hot spot corresponds to the star's magnetic axis. Because the magnetic axis is not aligned with the rotational axis, as the neutron star rotates the hot spot beams radiation into space like a searchlight, producing the observed pulses of radiation.

increases. A neutron star is thus like a bar magnet, of extreme strength, in space.

The details of how a fraction of the rotational energy of the star is converted into narrow pulses, as opposed to more diffuse radiation from around the star, are not very well understood, but some general statements can be made. Associated with the powerful magnetic field should be a strong electric field, which rips charged particles from the crust of the star. These particles are trapped in the magnetic field and forced to accelerate to high, perhaps relativistic, velocities. The photons that we receive as pulses are likely emitted from the regions around the magnetic poles of the neutron star. In general, the magnetic poles need not be aligned with the rotation axis of the neutron star. (This is hardly unusual; the magnetic axis of the Earth is misaligned with its rotation axis.) If the magnetic and rotation axes were coincident, we would receive a constant beam of radiation, and that only if our line of sight happened to look along the axis. However, if the emission comes from the magnetic poles, and these poles do not line up with the rotational poles, then the rotation will carry the beam around, sweeping it into our line of sight once per rotation. Only if the searchlight is pointed at an angle to its rotation axis can the lighthouse send a beam around the cape.

Many neutron stars seem to be solitary. This is not surprising, as we might expect that the violence of a supernova explosion would tear apart a binary, liberating, or perhaps even destroying, any companion the progenitor might have had. But some of the most interesting neutron stars

are not alone. For example, two known pulsars apparently have planets. It seems unlikely that primordial planets would survive a supernova, so it may be that a stellar companion was obliterated in the blast, then recondensed into a disk and assembled itself into one or more planets around the neutron star. If this is what happens, such a system must be very bizarre—a former star reincarnated as a planet, orbiting the corpse of its erstwhile companion. Other neutron stars are members of normal binary systems. The dynamics of such a system are quite similar to that of a white-dwarf binary, with some interesting twists due to the presence of the neutron star. The accretion disk around a neutron star would be much hotter and more energetic than that around a white dwarf. Gas piling up on the surface of a neutron star would find the crust to be even more unyielding than that of a white dwarf, and repeated episodes of sudden thermonuclear fusion might occur. Such a model explains the *X-ray bursters*, sources that emit spurts of X-rays at irregular intervals. Most such bursters are located near the center of the Galaxy or deep within dense clusters, environments where the density of stars is fairly high and thus where a significant population of neutron stars could be expected to have formed.

An even more bizarre effect can occur in the vicinity of a neutron star. Gas spiraling toward the rapidly rotating neutron star could be flung out at relativistic velocities in two *jets* collimated along the axis of rotation. The enigmatic object SS433, a star system located approximately 16,000 light-years from Earth, might be an example of such a system. The spectrum of SS433 reveals a mixture of approaching and receding gas with unusually large Doppler shifts; the spectra also show a smaller, regular shifting with a period of 164 days. The Doppler shifts indicate that gas flows at up to a quarter of the speed of light. The best explanation for this object is that it is a binary, one of whose members is a star that has overflowed its Roche lobe; the invisible companion is probably a neutron star. The strong emission lines emanate from a pair of relativistic jets emerging from the neutron star, one directed toward our line of sight (approaching) and the other away from it (receding). The regular shifting occurs because of *precession*, a wobbling of the axis of rotation due to gravitational torques upon the neutron star. SS433 exhibits on a small scale behavior similar to that seen in the cores of a fraction of galaxies, and especially in the cosmological objects known as *quasars*. Accretion around a neutron star, or around an even more dense object at the center of a galaxy, may be a common phenomenon throughout the universe.

SS433: the star that is “both coming and going”

Interactions with other stars can affect neutron stars in other ways. Although most pulsars have periods of a few tenths of a second, in the 1980s a new class of pulsars that spin with a mind-boggling period of a few thousandths of a second was discovered. These *millisecond pulsars* are thought to be the product of mass transfer from a companion. The accreting matter would be rotating, so as it struck the surface of the neutron star it could add angular momentum to it, thus increasing the neutron star’s rotation rate. Neutron stars are so small and so rigid that

they can spin with amazing periods, but even so the millisecond pulsars are probably near the breakup limit. We currently know of over 100 millisecond pulsars, roughly half of which are in globular clusters and half in the main part of the Galaxy. Many of these are in binary systems, but some are solitary. A solitary millisecond pulsar was probably ejected from the binary by an encounter with another object.

Even stranger are the double pulsars. If it is unlikely that a binary system could survive one supernova, it seems nearly impossible for it to survive *two*. Yet a handful of binary neutron star systems have been discovered. Perhaps both progenitor stars lost quite a lot of mass prior to exploding, so their supernovae were not excessively violent. Alternatively, perhaps the two pulsars did not form together. A solitary pulsar might have interacted with an existing binary, displacing the normal star. Whatever the formation mechanism, the binary pulsar is a fascinating and important object. The first binary pulsar discovered provided the first firm, albeit indirect, evidence for the existence of gravitational radiation, waves in the fabric of space-time itself. Just as the emission of electromagnetic waves causes the emitter to lose energy, so do gravitational waves carry away energy, causing the objects' mutual orbit to decay gradually. Pulsars are highly accurate clocks, making it possible to measure with great precision the slow decrease in the orbital period as the system radiates gravitational waves.

The white dwarf is supported by electron degeneracy pressure and has an upper limit on its mass. The neutron star is supported by neutron degeneracy pressure, and it too has an upper limit to the mass that can be so supported. Astrophysicists are not entirely certain what that upper bound is; the physical state of matter at these extreme densities and pressures is not as well understood as we would like. However, the limit almost certainly lies between two and three times the mass of the Sun. If an imploding stellar remnant finds itself with more than this mass, this unfortunate star cannot halt its collapse as a neutron star. Modern physics knows of no force sufficient to prevail against gravity, and the star collapses to a *black hole*. The black hole has properties so strange that we cannot appreciate them until we have made a more careful study of the structure of the universe.

Black holes: the end point for the most massive stars

Chapter Summary

Stars play several important roles in cosmology. Stars make up the majority of the luminous matter in the universe, and many cosmological questions are related to the lives of the stars. How many stars are there? What are their masses? How much of the mass of the universe is made up of stars, including those too dim to see? How are stars born? How do they die?

Stars are born in huge clouds of interstellar gas and dust. In the hearts of these molecular clouds, gravitational forces overwhelm regions of cold gas, drawing these cold clumps into dense cores. From such cores stars will form. Once formed, stars fuse hydrogen in their cores. Stars that burn hydrogen are called main sequence stars, from their locations on the Hertzsprung–Russell diagram,

and this phase occupies the majority of a star's life. A star's luminosity increases as the third power of its mass. Since a star's lifetime will be affected mainly by its mass divided by its luminosity, it follows that the more massive stars die earlier. The main sequence can be used to determine the ages of the oldest stars in a cluster. The most ancient star clusters are the globular clusters. Determining the ages of stars in such clusters provides a lower limit to the age of the universe. Currently the best data and calculations imply that the oldest clusters are 12 to 15 billion years old.

Astronomers know that there is a great deal of mass in the universe that is not in ordinary stars. This unseen material is often called the *missing mass* or *dark matter*. Among the possible candidates for this dark matter are small star-like objects that have too little mass for nuclear fusion to occur in their cores. These objects are called brown dwarfs. Other candidates include the remnants left after stellar death. A star begins to die when it runs out of usable hydrogen fuel in its core. The next most easily fused element is helium. When the star begins to burn helium at its core, its outer layers swell drastically, causing the star's size to expand enormously and creating a *red giant*. The helium-burning red giant phase of the

star's existence is relatively brief. Stars with masses of up to about 6 times that of the Sun expel their outer layers; the remnant core collapses until the electrons of its atoms can no longer be squeezed any closer. This phenomenon, called electron degeneracy, is a consequence of the Pauli exclusion principle of quantum mechanics. The core, now a white dwarf, continues to shine feebly as light diffuses through it, cooling over billions of years until finally it leaves behind a dead, compact, black dwarf.

More massive stars are able to fuse elements heavier than helium, up to iron; at this point, no more energy production is possible. If the star was able to shed enough mass during its giant phase, it might fade away as a white dwarf. If not, it collapses catastrophically, blowing its outer layers into space in a supernova. The core left behind is too massive even for electron degeneracy to support it; instead the electrons and protons are squeezed together into neutrons, and the core becomes a neutron star. Neutron stars are visible only when they beam radiation as they rotate, in which case we detect them as pulsars. If the core is too massive for neutron degeneracy to support it, however, it ends in the ultimate product of gravitational collapse, a black hole.

Key Term Definitions

interstellar medium Gas, dust, bits of ice, etc. that fill the space between the stars. Nearly all of the interstellar medium is hydrogen and helium gas, with hydrogen most abundant.

nebula A cloud of gas in space.

brown dwarf A substellar object that is near, but below the minimum mass for nuclear fusion reactions to occur in its core.

hydrostatic equilibrium The balance between gravity and gas pressure in an object such as a star.

ideal gas A gas in which the mutual interactions of the gas particles are negligible, except for their momentary collisions. The pressure is determined by the ideal gas law, which is the formula that relates temperature, pressure, and volume for an ideal gas.

deuterium An isotope of hydrogen whose nucleus contains one proton and one neutron.

primordial element Those elements and isotopes formed in the big bang; specifically, hydrogen,

helium (both helium-3 and helium-4), most deuterium and tritium, and some lithium-7.

main sequence The curve on a Hertzsprung–Russell diagram along which stable hydrogen-fusing stars lie.

Population I, II, III Labels for the generations of stars, determined by the proportion of heavy elements contained in their members. Population I stars are youngest, while Population III represents the primordial stars.

metal In astronomy, all elements heavier than helium, regardless of whether they are chemically “metals” or not.

globular cluster An aggregation of approximately 100,000 stars. Halos of globular clusters orbit many galaxies. Some globular clusters are thought to be among the oldest structures in the universe.

turnoff mass The mass of the largest star in a cluster that is still on the main sequence. The age at which a star moves from the main sequence to the red giant phase depends almost entirely upon its mass

and chemical composition, with more massive stars leaving the main sequence earlier. The stars in a cluster all formed at essentially the same time and have similar chemical composition, so the turnoff mass can be used to determine the age of the cluster.

red giant A star near the end of its life; it fuses heavier elements in its core and has a greatly expanded outer layer.

electron degeneracy A condition of matter in which all quantum states available to the electrons are filled.

white dwarf A compact stellar remnant supported by electron degeneracy pressure and shining only by the diffusion of light from its interior. White dwarfs cool slowly; if the universe exists long enough they will all cool into nonluminous black dwarfs.

accretion disk A disk of gas that accumulates around a center of gravitational attraction, such as a white dwarf, neutron star, or black hole. As the gas spirals in, it becomes hot and emits light or even X-radiation.

nova An abrupt, very bright flare-up of a star. Most likely due to the accumulation of hydrogen from a companion star upon the surface of a white dwarf. The pressure and temperature grow in the deposited matter until a thermonuclear explosion is generated.

Chandrasekhar limit The maximum mass, approximately $1.4 M_{\odot}$, above which an object cannot support itself by electron degeneracy pressure; hence it is the maximum mass of a white dwarf.

supernova The explosive death of a star. Type Ia supernovae probably occur when a white dwarf accumulates upon its surface too much gas from a companion, causing the white dwarf to exceed the Chandrasekhar limit. Type II supernovae occur when a massive star has reached the end point of nuclear fusion and can no longer support itself. In both cases, the result is a catastrophic gravitational collapse and an explosion so violent that elements heavier than iron are created. Any remaining core becomes a neutron star or a black hole.

neutron degeneracy A condition of matter in which electrons and protons are crushed together to form neutrons, and all quantum states available to the neutrons are filled.

neutron star A dead “star” supported by neutron degeneracy pressure.

pulsar A rotating neutron star that emits regular, periodic bursts of radio emissions.

conservation of angular momentum The principle that the angular momentum of a system, the momentum of rotation about a point, remains the same as long as no external torque acts.

Review Questions

- (5.1) What objects in the Galaxy are the most likely stellar nurseries? What properties make them good locations for star formation?
- (5.2) Distinguish between brown dwarfs and red dwarfs. Are brown dwarfs common? What would be the significance of a huge number of brown dwarf stars?
- (5.3) What is hydrostatic equilibrium, and why is it important to the existence of stars?
- (5.4) What are the main physical characteristics that control the life of a star?
- (5.5) What is the main sequence? What is the significance of the main sequence turnoff point in a cluster of stars, and how can this turnoff mass be used to obtain an estimate of the cluster’s age?
- (5.6) What happens in a nova? How does it differ from a supernova? Some science fiction stories have used plots in which the Sun threatens to explode as a nova or a supernova. Is this a possible scenario?
- (5.7) Why is there an upper limit to the mass that can be supported by electron degeneracy pressure?
- (5.8) Explain why a Type Ia supernova makes a better indicator of distance than a Type II supernova. Why does a supernova make a good distance indicator for cosmology as compared to ordinary stars?
- (5.9) Describe three ways in which the study of stars can provide important cosmological information.

- (5.10) Suppose a stellar core with a radius of 30,000 km rotates once, that is, 2π radians, every 5.2×10^7 seconds (about 25 days). Let the mass of the core itself be M_c . The star undergoes a supernova and the core collapses to a neutron star. Assume that no mass is lost from the core (an unrealistic as-

sumption, but adequate for this example), but the radius decreases to 30 km. Assume that both the progenitor and the neutron star are approximately spherical. What is the new rotation rate of the star?

This page intentionally left blank

Part III

Relativity

This page intentionally left blank

Infinite Space and Absolute Time

6

Nothing exists except atoms and empty space; everything else is opinion.

Democritus (460–370 BCE)

Creating the universe

What is the universe? If the universe is everything, can there be anything beyond it? Where do we fit into the universe? How was the universe created? What will be its eventual fate? With all the complexity that we see immediately around us, how can we hope to understand something so intricate on even larger scales? Such questions have been asked for as long as we have any traces of human thoughts. The answers that have been imagined have been profound, or philosophical, or fanciful, or stern; but until the development of modern science, the explanations offered had less to do with the way the universe was, than the way humans imagined it might be. The universe remained mysterious and ineffable. This slowly began to change with the ancient Greeks, who saw a universe built on geometry, a universe that was just as beautiful in its mathematical harmony as any mythological cosmology. With the development of Newtonian mechanics, the universe began to seem comprehensible to the human mind. Modern science has brought about the development of models of the universe that can be compared with and tested against observation. These models incorporate the inferred natural laws that give coherence to our observations, and enable us to predict previously unobserved phenomena. We may comprehend no reason that the real universe must obey any laws at all, particularly those of human construction, but we can say with confidence that our rules describe *something* about the real universe.

The universe that is accessible to science is the *physical* universe; the universe of material objects, of energy, of space, and of time. This universe contains all that is physical, including all things that are observed, anything that affects or influences other observables, all that is affected by physical things, and hence everything that is subject to experiment and scientific proof, or disproof. Atoms, particles, energy, forces, the laws of nature, even space and time, are physical. Everything composed

Key Terms:

- anthropic principle
- Copernican principle
- isotropy
- homogeneity
- cosmological principle
- perfect cosmological principle
- coordinates
- velocity
- speed
- acceleration
- inertia
- mass
- inertial motion
- frame of reference
- inertial reference frame
- inertial force
- inertial observer
- invariance
- relativity
- Galilean relativity
- luminiferous ether
- interferometer

The physical universe

What is time?

of matter, or subject to the laws of nature, must also be physical and hence part of the universe. Anything that is not part of the universe cannot, by definition, have any physical properties. This definition keeps our cosmological considerations meaningful and consistent.

It might seem obvious to regard matter, energy, physical laws, forces, and the like, as physical things, but the inclusion of space and time in this list requires further justification. Time as a physical quantity seems especially troubling to some, since it appears to be at odds with much human experience. The rate of the passage of time can seem to vary depending upon one's mood; a pleasant day may fly past, while an unpleasant hour may seem to last forever. Time might even seem to be a human artifice. Yet this clearly cannot be true. Human perceptions of a quantity are distinct from that quantity. The human brain is capable of keeping track of short time intervals with impressive accuracy, but it can be easily fooled if distracted or bored. This is actually just as true of space as of time. Many well-known illusions depend upon tricking the systems in the brain that estimate distance intervals or relative sizes; yet space often seems more concrete than time. Moreover, the conceit of time as a human construction smacks of anthropocentrism. There is clear evidence that the universe has changed, that it has a history; but most of this history, not only of the universe but even of the Earth, has passed without the presence of humans. Thus time must have existed before humans came into being. Furthermore, the universe is very much larger than the sphere of human influence, yet periodic physical processes clearly occur in all parts of the universe, so time must exist where there are no humans. Time and space play a role in the laws of nature independent of humans. The issue that has faced scientists is *how* time and space enter into the construction of the universe.

Philosophers have debated through the centuries whether or not space and time can be said to exist in their own right, or whether they are only relations between physical things, where "things" can make the sole claim to existence. The modern theories of special and general relativity make it quite clear that space and time *are* physical; they can influence matter and energy and, in turn, be affected by matter and energy. They are active participants in the history of the universe. However, the inclusion of space and time as physical components of the universe has certain consequences. Any model of the universe must include and explain space and time along with every other physical phenomenon. Indeed, it is possible to create models of the universe that contain space and time alone, yet still change and evolve. Thus it is not permissible to invoke a pre-existing space and time in which to construct the universe. For example, it is not meaningful to ask "what happened before the universe existed?" or "what is outside of the universe?" because both of these questions assume the existence of attributes ("before" and "outside") that must posit space and time as properties distinct from the universe itself. Yet time did not exist before the universe, and space does not exist outside it. The big bang did not happen "somewhere."

The universe is not expanding into space nor even into space-time. Do not think of the universe as embedded in something larger.

The confusion over the physical nature of space and time carries over into one of the thorniest cosmological questions: the creation of the universe. When humans ponder the creation of the universe, generally the question they ask is, “Why does *something* exist rather than *nothing*?” Why is there a universe at all? In framing that question, the state of nothingness might be imagined as a great emptiness, extending in all directions and lasting an exceedingly long time. The flaw in this image is that time and space are physical entities, so empty space moving forward in time already describes something. How, then, were space and time created? Since we cannot help but imagine an act of creation, or, for that matter, any action, in terms of space and time, how can we contemplate some unknown metastate in which this ultimate act of creation occurred? This issue is sufficiently disturbing to some cosmologists that they attempt to sidestep it by extending the history of the universe into an indefinite, infinite past. If there is no point at which $t = 0$, the reasoning goes, there is no need for creation. However, the question of existence is not answered by supposing that the universe is infinitely old. Time is physical, and an infinite time would be just another physical attribute of the universe. Indeed, whether or not the universe has infinite extent in time is a question not much different from the superficially less disturbing issue of whether or not the universe is spatially infinite or finite. An infinitely old universe is not *nothing*, so it must have been created; it was simply created with time that extended infinitely, in the same way that the universe may have been created with infinite spatial extent.

Some relativists and cosmologists, most prominently Stephen Hawking, have pointed out that in general relativity, finite space and finite time can form a completely self-contained, finite *space-time* with no boundary or edge at all. The point we call $t = 0$ only appears to be a boundary in time because of the way in which we have divided space-time into space and time. Such a universe can be contemplated with the help of an analogy to the Earth. On the Earth, the North Pole is the limit to how far it is possible to travel in the direction we call north, but it is nevertheless just a point on a continuous, boundaryless globe. Similarly, the point $t = 0$ in a spherical big bang model of the universe represents merely an arbitrary demarcation in time. Without boundaries (spatial or temporal), there is no need to imagine the universe to be contained within some meta-universe. Like the infinitely old universe, the spherical universe attempts to avoid the question of creation by eliminating $t = 0$ as a special point in time. There may be any number of reasons to prefer a universe of infinite or finite time, infinite or finite space; there are certainly detectable differences among these types of models. But the presence (or absence) of a $t = 0$ point in time provides no answer to the mystery of creation, nor does it have implications for the existence of a creator, beyond those provided by the mere fact of existence. There is little, if anything, that can be said about

The ultimate question of creation

the metaphysical creation of the universe. Since our observations are of physical attributes, and science deals with physical things, the issue of creation, which must necessarily be metaphysical, cannot be addressed. The universe might be here because of the action of some creator, or maybe it “just so happened.” At present, it is not possible to ask this question in a way that is scientifically testable.

The initial conditions of a universe

In scientific cosmology, we confine our attention to well-posed questions, those we might be able to answer. We can ask questions such as: what is the universe like right now? How did it arrive at this state? Traditionally, we would answer such questions with a description of the observed universe and a statement of the laws of physics, laws that we believe describe the time history of the universe. If we trace the evolution of the universe backwards in time, we can ask whether or not there was a point $t = 0$. If there was, our exploration must eventually arrive at the question of *initial conditions*, the description of how things were at the earliest possible moment that we can contemplate. The science of cosmology aims to describe those initial conditions and to answer the question of how the universe evolved from them. There are many possible sets of initial conditions, and we must adopt criteria for what we shall hold as good initial conditions. As an example, suppose we were to assert that the universe was created at 7:20 this morning. In such a case, everything we know must have been created from nothing at that moment, including the stores of memories in our minds, light arriving from distant stars at the Earth, fossil bones in the ground, and history books describing a past. This is clearly a very complicated set of initial conditions. Moreover, such a model cannot be disproved, because any condition one might propose as a test could simply be lumped into the initial state that was created at 7:20 am. This lack of testability means that such a model fails as a scientific theory. If we compare the initial conditions in the “7:20” model with the big bang initial conditions, we find that in the big bang model, the universe began in a much simpler state. There was a certain amount of energy and matter, certain physical laws, and certain fundamental constants. The complexity of the universe we observe existed as a potentiality, and developed naturally in the subsequent evolution. The big bang model is testable and falsifiable because the initial conditions are constrained by the laws of physics. The theory makes specific predictions as to what the early universe should look like and how it should subsequently evolve.

In formulating our cosmological models, we would like to be able to describe the initial state of the universe in as few terms as possible. In science we generally adhere to the principle of Occam’s Razor; in the absence of compelling evidence to the contrary, the simplest of competing explanations is preferred. The big bang universe has the virtue of relative simplicity of its initial conditions. As our understanding advances and theory approaches ever nearer to $t = 0$, the initial conditions of the big bang seem to become even simpler. Yet even with the comparatively simple set of initial conditions afforded by the big bang model, there are interesting and challenging questions to consider. For example, the fun-

damental constants of nature, such as the gravitational constant G , the speed of light c , and Planck's constant h , are held by current physical theories to be constant in space and time, and hence part of the initial conditions. The particular values of the fundamental constants, along with the basic laws of physics, determine what is permitted in the universe. If any of these conditions were changed, even slightly, then the universe that would result might be quite different from the one we observe. What if nuclear reactions were not possible at the densities and temperatures prevailing in the cores of gravitationally bound conglomerations of gas? Would there still be stars? What if chemical constants were sufficiently altered that carbon could not form the long chains found in organic molecules? In either such hypothetical situation, or in many others, life, as we understand it, might not develop.

We do not know why the fundamental constants have the values they do, or whether they could have taken other values. But we can imagine that all things were possible and, out of all potential universes, ours is special by dint of our presence in it. The fact that our existence carries implications for the nature of the universe is known as the **anthropic principle**. Its most basic form, the *weak anthropic principle*, states that the conditions we observe in the universe must be compatible with our own existence. The weak anthropic principle sifts out all conceivable universe models that do not admit the possibility of the development of life. The cosmologist Fred Hoyle is said to have invoked the weak anthropic principle to predict the existence of an excited state of the carbon atom, since such a state allows the triple-alpha nuclear reaction to create carbon in stars.¹ Since Earthly life depends on the existence of carbon atoms, we can infer that the necessary excited state must exist. Many find the anthropic principle appealing because it appears to give a special role to our existence, but in fact it says nothing inherently more profound about life *per se* than it says about atoms, or stars, or galaxies. In the example above, the mere existence of carbon is sufficient; the carbon has no compunction to form a basis for life. In a universe with different physical conditions, carbon may still have been able to form by some other means, or else life might be based on another atom. By itself, the weak anthropic principle is not even really a testable scientific hypothesis; it is merely a restatement of the requirement that our models be consistent with observation.

The anthropic principle

A more stringent, and controversial, form of the anthropic principle, the *strong anthropic principle*, states that the initial conditions occurred *because* we are here; that is, our presence here and now somehow affected the initial conditions such that we could eventually arise. Thus according to the strong anthropic principle, the conditions necessarily existed so that we can exist; the purpose of the universe is to create life. The strong anthropic principle does not explicate how this backwards influence might have been exerted, but does seem to require forethought on

¹Chapter 5 discusses this stage of stellar evolution.

The apparent specialness of the universe

the part of the universe. This takes it beyond the bounds of science and into teleology, the attribution of intent to the universe as a whole.²

Some people are drawn to the strong anthropic principle because it asserts a meaning to the universe, and that meaning is us. To the student of history, however, this is very familiar. As we have seen, most myths included a central role for humans. In the absence of any scientific basis for the strong anthropic principle, we again enter the realm of mythology. It is sometimes argued that even though we do not have any basis for the strong principle, the fact that we are here, and the apparent specialness of the universe, must be telling us something. This is possible, but there is a weakness in this position: we have no grounds for concluding that this universe is really so special. We have but one example; we have no way of knowing what might be possible, or what the alternatives might mean. As an illustrative example, consider what might have happened if your father had been killed in a war before you were conceived. If that had occurred, the you that exists here and now would not exist to ask such a question. Hence your very existence necessarily (and tautologically) implies that your father lived at least long enough for you to be conceived. But it does not imply that the purpose of your father was to produce you, and hence the war's outcome was pre-ordained. In this case it was a matter of chance. Things happened as they happened. Each of us is here by a happy accident of conditions.

Throughout the history of cosmological thought, humanity has struggled to realize that our planet, our star, and our galaxy were not unique, but merely individual members of a far greater collection of planets, stars, and galaxies. The wonderful appropriateness of the conditions on the Earth to sustain us, and the suitability of the Sun to warm us to just the right temperature, and to contain sufficient nuclear fuel to last long enough for us to evolve and come into being, might seem to be conditions that would be extraordinarily rare. Yet the discovery of myriads of other stars and galaxies, and recently of extrasolar planets, implies that there must be many other planets throughout space: some too large, too small, too hot, or too cold for life to develop. Similarly, there are many stars too bright or too dim, or whose lives were too brief, to nurture life. Could this principle be extended to the universe itself? Perhaps this universe that seems so special is not the only one of its kind. As we ponder a universe that provides the conditions necessary to bring us into existence, we might conclude that there are other universes, perhaps in infinite numbers, that are less hospitable. From contemplation of the weak anthropic principle arises the possibility of multiple universes.

There may be multiple universes. It may be that the one and only universe must contain life, and that the initial conditions had to be what they were. It is equally conceivable that it “just so happened.” For the moment we have no scientific basis for any conclusions. The whys of

²If the intent of the universe is to create life, then it has done so in a very inefficient manner. For example, Aristotle's cosmos would satisfy the strong anthropic principle, and would give a much greater amount of life per cubic centimeter to boot!

creation remain a mystery. But describing the subsequent unfolding of that creation will prove challenge enough.

The cosmological principle

The nature of time and space have always been at the heart of humanity's cosmological musings. Early anthropocentric cosmologies placed humans at the center of the All, creating in the process a very special attribute of space: a center. Similarly, creation stories tended to place specific restrictions on time, such that the history of the universe coincided more or less exactly with that of humankind. The geocentric universe of Aristotle was more physical than the earlier anthropomorphic mythologies, but it still placed the Earth, the home of humankind, at the spatial center. Aristarchus, and later Copernicus, moved the center of the universe from the Earth to the Sun, the first significant loss of status for humanity. The Sun-centered, or heliocentric, view is correct for our solar system; the Sun *is* at the center of motion of the planets. But what about the universe as a whole? Is there a center to the universe? The center of the universe, if it exists, must be a special place, if for no other reason than that it is unique. But the universe is, in virtually every model since Newton, a very large place. What are the chances that the solar system would occupy such a special location? Essentially zero, of course. Observations have progressively demonstrated that the Earth is not the center of the solar system, that the Sun is not the center of the Milky Way Galaxy, and that our galaxy is not the biggest we can see, nor is it even at the center of its modest cluster of galaxies. While we cannot decisively prove that we do not lie near some center of the universe, the history of human cosmological thought suggests that a certain humility is in order. The principle that the Earth or the solar system does not occupy any special place in the universe is usually called the **Copernican principle**. This principle does not claim that no center exists; only that we are not located there. Even if we accept that we are not at the center of the universe, might there yet be a center somewhere? Since we cannot see all of the universe, we are unable to answer this question from direct knowledge. Instead, we must bring to bear certain concepts that will aid us in understanding the overall structure of the cosmos. Two very important such concepts are *isotropy* and *homogeneity*.

Does the universe have a center?

The Copernican principle

Isotropy is the property of uniformity in all directions. No single direction is special or distinct from any other. One example could be a forest of indefinite size, with identical trees and level terrain as far as the eye can see, regardless of direction; nothing enables any particular direction to be distinguished from any other. Such a forest is isotropic. Now suppose a trail cuts through this forest. The forest is no longer isotropic; the trail selects a preferred direction. The surface of an unmarked sphere provides another example. All directions are equivalent; the sphere is isotropic. Contrast this with the surface of a cylinder. A

The isotropic universe

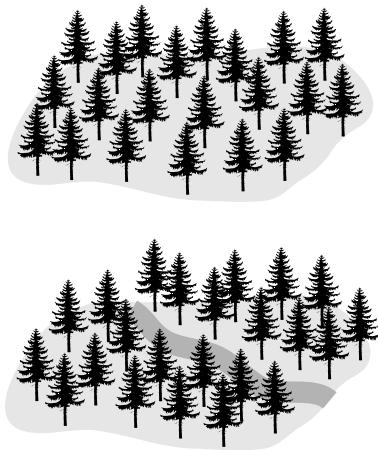


Fig. 6.1 (top) An isotropic and homogeneous forest looks the same in all directions. **(bottom)** A homogeneous but anisotropic forest has preferred directions selected by a system of trails, but on sufficiently large scales looks more or less the same everywhere.

The homogeneous universe

cylinder has a long direction parallel to the axis, and a short direction around the axis. The cylinder is not isotropic.

To a certain extent, we can test the universe for isotropy. We need merely make observations in all spatial directions and determine whether there is any systematic trend, or dependence upon direction, for any measurable property. We can define special directions in space, such as the directions toward the Sun or the Galactic center, but these are strictly local properties, rather than universal attributes. To test for isotropy, or anisotropy, of the universe as a whole we must examine the largest scales, such as the overall distribution of all observable galaxies, or the distribution of quasars throughout the sky. As far as we can tell from the observations, the universe is indeed isotropic at the largest scales. Such measurements are prone to various observational errors, however, not the least of which is the fact that the most distant galaxies are the most difficult to see. Thus we cannot unequivocally declare from these indications that our universe is isotropic, although isotropy remains the most viable, as well as the simplest, interpretation of the distribution data. The strongest evidence for the large-scale isotropy of the universe is the cosmic background radiation. This background radiation consists of microwave energy that is present at every point in the sky and has the spectral distribution of a blackbody, at a temperature of 2.725 K. The best explanation for this radiation is that it is the afterglow of the big bang. It is observed to have very nearly equal strength and temperature in all directions, after we account for the motion of the Earth. The uniformity of this cosmic relict constitutes an important testimony for the isotropy of the universe.

The second concept that will aid us in our quest for the structure of the universe is **homogeneity**, the property of similarity of all locations. The surface of an unmarked sphere is homogeneous: every point is the same as every other point. The surface of a cube is not homogeneous: the edge points are different from the points on the cube faces. A dense forest can seem quite homogeneous, with few detectable differences on the average, over many square miles. If the universe is homogeneous, then all points throughout all space are more or less equivalent, and everywhere the same physical laws are obeyed. It is difficult, perhaps even impossible, to test the universe for homogeneity. We cannot visit, or even see, all possible points in the universe. But from what we can see, it looks fairly homogeneous. Distant stars and galaxies resemble nearby stars and galaxies. The same elements we find on Earth are present in the farthest quasars.

It is possible for a universe to be isotropic but not homogeneous; however, this occurs only in the special, and rather contrived, case that some central point exists, and isotropy holds only at that single point. An example of an isotropic, but inhomogeneous, situation would be the pinnacle of the only hill in a huge forest. From this point the scenery would look the same in every direction. But the observed isotropy holds only at the peak of the hill. Away from the peak there would always be a special direction: upward to the summit of the hill. Hence if the

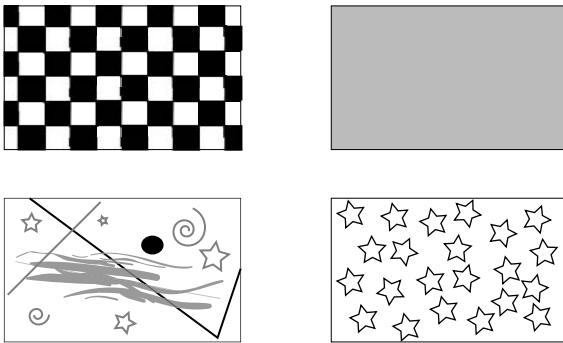


Fig. 6.2 Which representative samples of two-dimensional universes are homogeneous, isotropic, or both?

universe seems to be isotropic, that is, the same in all directions, then either it really is the same everywhere, or else we live at a unique point where the universe gives the appearance of isotropy. Therefore, despite our inability to examine all of space, we can infer that the universe is probably homogeneous by noting that it appears to be isotropic on the largest scales. If we apply the Copernican principle to state that we are not at a special location, then the universe must look more or less isotropic to *all* observers and must, therefore, be homogeneous. Isotropy plus the Copernican principle implies homogeneity.

It must be emphasized that while isotropy can imply homogeneity, the converse is not true. Any universe that is isotropic and has no special point must be homogeneous; whereas the universe could be homogeneous but not isotropic. Isotropy demands that there be no preferred direction, whereas homogeneity merely requires that the universe have the same appearance everywhere. A forest with a trail cannot be isotropic, since the trail clearly defines a special direction. If there is only one trail, then it would also delineate a set of special locations, so neither would this forest be homogeneous. But suppose that there is a network of trails running north and south, cut through the forest every kilometer. This forest would be homogeneous on large scales, but not isotropic. Geometrical figures provide other examples: the surface of an infinitely long, uniform cylinder is homogeneous but not isotropic, because there are distinguishable directions, along the axis and around it. A spherical surface, on the other hand, is both homogeneous and isotropic.

Figure 6.2 shows sections of two-dimensional universes. These sections are simply representative, and the universes actually extend indefinitely. For example, the section that is a bland, uniform gray is both homogeneous and isotropic; every point is the same, and every direction looks the same. The checkerboard can be considered homogeneous on a sufficiently large scale. While there are local variations (black and white squares), these same patterns appear everywhere. There is a sense of direction, however. On the checkerboard it is possible to proceed along the alternating colors of the squares or in a diagonal direction along squares of one color. These directions are quite distinct.

$$\text{Isotropy} + \text{Copernican principle} = \text{homogeneity}$$

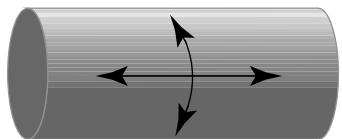


Fig. 6.3 The surface of a uniform cylinder is homogeneous but not isotropic.

The cosmological principle states that the universe is homogeneous and isotropic

The same physical laws apply throughout the universe

The concepts of isotropy and homogeneity of the universe are combined into one overall principle, the **cosmological principle**, which asserts that all points and directions in the universe are more or less equivalent, and thus that the universe is both homogeneous and isotropic. Given the cosmological principle, we conclude that there is no center of the universe. All points in space are basically equivalent; there is no single point that is central, or in any other way special. The adoption of the cosmological principle completes the process begun by Copernicus. Not only is the Earth not the center of the universe, there is no center at all. It is not surprising that the cosmological principle came rather late in the history of humanity's thinking. Seen from the surface of the Earth, the universe appears to be anything but isotropic. The stars are not distributed evenly, but are concentrated in a broad band, the Milky Way, which stretches across the sky, delineating a direction. The uniformity of brightness in this band led astronomers at first to conclude, incorrectly, that the Sun was near the center of a great disk. Only later was it realized that appearances are deceiving, and we are actually closer to the edge of the disk of the Galaxy. Astronomers formerly assumed that the Milky Way constituted the bulk of the universe, but improvements in telescope technology laid that fallacy to rest. In our universe we see galaxies, and clusters of galaxies, and clusters of clusters of galaxies, all containing galaxies of different sizes and shapes, for as far as we can detect their light. The Milky Way is nothing unique, after all. On what scale does the universe become truly homogeneous? Even now that question is as yet unanswered, but it does appear that on the largest scales, those most suitable for cosmology, the universe is isotropic, and, by implication, homogeneous.

The cosmological principle goes far beyond a simple assertion that the universe has the same appearance everywhere, to include all physical properties. Only by an appeal to the cosmological principle can we posit that the same laws of physics discovered on Earth also apply to distant galaxies, and that all objects, no matter how far from us, are composed of the same fundamental substances as we find on the Earth and in its vicinity. This is clearly a sweeping generalization that might seem to reach beyond our capabilities; but without something like the cosmological principle, how could we ever hope to understand anything about our universe? In fact, it can be argued that the most important aspects of the cosmological principle relate to the uniformity of physical laws. We might easily imagine a universe that was not isotropic in its distribution of matter, even at very large scales; such models have been advanced, even quite recently. But we could not hope to understand a universe in which physical laws varied willy-nilly from one region to another. If the same spectrum originated from different elements, or indeed if the elements themselves had different properties elsewhere, we could say precious little about distant galaxies or quasars. The cosmological principle is an assumption about the nature of the universe. Like all scientific postulates, it is unprovable. It is, however, disprovable;

its continued justification depends on the coherence and success of the models that utilize it.

If there are no special directions or locations to space, what about time? The cosmological principle asserts that the universe is homogeneous, but such a universe need not be static or unchanging; it requires only that at a given time, all points must appear the same. Using the forest analogy again, we might begin with an empty clearcut onto which Douglas-fir seedlings are planted uniformly. The trees grow at roughly equal rates, and at any given time the forest looks the same, but it still changes with time. There is a more restrictive principle that holds that there is no special point in time, as well as in space. This is known as the **perfect cosmological principle**, and it states that the universe is the same at every point in space and at every point in *time*. Continuing with the forest analogy, an old-growth mixed forest would not be homogeneous in the same way as a stand of Douglas fir, but it would be difficult to distinguish one location from another over a fairly large scale. One patch of forest would have about the same number of redwood trees, fir, spruce, pine, etc., as would any other patch. One of the properties of old-growth forests is that once established, they reach an equilibrium in which new trees grow at exactly the rate needed to replace those that die; the age of the forest would be as indeterminate as a location within it. By the same type of argument, any universe that obeys the perfect cosmological principle must appear to be the same, on the average, everywhere and for all times. Such a principle is extremely restrictive. Indeed, the perfect cosmological principle goes too far, and has been disproved. Observations indicate that the universe *does* have many special points in time, and does evolve with time.

The perfect cosmological principle predicts a steady state universe

Cosmological models are intimately linked to the philosophy behind the physical laws held to govern the universe. The cosmological principle is one possible paradigm. Before the advent of the modern model, other physical theories informed other cosmologies. The interdependence of cosmology and physical philosophy is sufficiently great that the failure of one could bring down the other as well. From the age of the Greek philosophers until the present, cosmology and physics have advanced, or declined, hand in hand. The Aristotelian universe is an example that is clearly neither homogeneous nor isotropic in space, not only in its appearance but also in its physical laws. According to Aristotelian physics, Earthly objects moved through space linearly, toward the location that was appropriate to their percentages of earth, fire, water, and air, while celestial motions were perfect circles executed forever. The Aristotelian cosmology was in accord with Aristotelian physics. Special points and directions were inherent to the model. Space was defined only in terms of the objects it contained; Aristotle could not conceive of the vacuum of space, and stuffed his model with tangible physical entities. Not only the Earth and heavenly bodies were physical, but also the spheres that bore the planets and stars on their daily travels had real physical existences. Aristotle would have denied any possibility of travel to the Moon, for the traveler would be unable to continue with

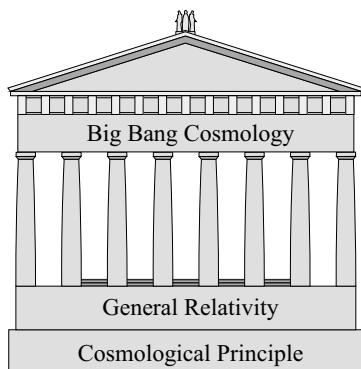


Fig. 6.4 Modern big bang cosmology is built on the foundation of general relativity and the cosmological principle.

linear motion in the celestial realm, and would probably smash into the Moon's crystalline sphere as well. On the other hand, the Aristotelian model was, more or less, unchanging in time. Aristotle's concept of time seems to have been rather ill defined, but it functioned as a marker of occurrences. Even here, however, the inhomogeneity in space played a role; change occurred only on Earth, not in the heavens.

Newtonian physics, in contrast, makes no special distinctions in space or in time. Newton's laws of motion contain no preferred directions, nor does location have any inherent effect upon mechanics. Newtonian physics depends implicitly upon the existence of an absolute space and time to which motion is referred. Whether an acceleration is present can be determined by measuring the change in velocity with respect to markers laid down in absolute space and time. The markers themselves, which might consist of the background of fixed stars, or any other appropriate standard, are merely convenient references that have no intrinsic significance of their own. Space and time have an independent existence, regardless of how we choose to measure them. Newton's cosmology reflected his mechanics. The universe consisted of stars scattered about uniformly everywhere in space; the stars either lived forever, or died and were recreated. This grand machine was set into motion at some specific point in time, but throughout its existence, the universe looked the same for all locations and all times. Newtonian physics was everywhere valid, and a knowledge of the initial conditions would, in principle, enable a perfect computer to calculate the entire destiny of the universe.

Just as the claustrophobic and rather judgemental Aristotelian-medieval universe troubled some thinkers of its time, so did Newton's aggressively deterministic cosmos create doubts among many philosophers of the Enlightenment. Not only did it seem to preclude any free will on the part of humans, but it made some strangely rigid assumptions. One difficulty was that Newton's law of gravity required a force to act instantaneously across empty space. What conducts that force? Absolute space and time, which affected everything but which were affected by nothing, were also particularly repugnant to some scientists of the day. Moreover, it was recognized even then, and by Newton himself, that the Newtonian universe depended on a very delicate balance; since gravity is strictly attractive, its force would inexorably pull lumps of matter together. The only way to prevent the Newtonian clockwork from collapsing onto itself was to assume an infinite, perfectly uniform distribution of matter. Despite these background rumblings, however, Newtonian mechanics was an indisputable success, and the weaknesses of the corresponding cosmological model were swept under the rug for two centuries. After all, it had no compelling competitors at the time.

Taking measurements

Either this man is dead, or my watch has stopped.

Groucho Marx

The modern viewpoint that arose during the 20th century flows from and around the cosmological principle; to understand modern cosmology, we must explore its relation to the form of modern physical theory. This journey will take us from grand galaxies to the elementary particles, but underlying all of it will be the meaning of space and time. Let us begin, then, by contemplating how we can quantify the relationships among space, time, and our observations of the universe. The scientific revolution introduced the importance of measurements into our conceptions of the universe. Pure thought alone cannot reveal the nature of the universe any more than it can manufacture gold. Lord Kelvin expressed the issue quite succinctly when he remarked that we cannot truly understand anything we cannot measure. Careful measurement is fundamental to the attainment of scientific knowledge through scientific *observations*. We must measure physical properties in a repeatable manner that is unaffected by the observer or by the instrument. We all may feel intuitively that spatial and temporal relationships exist between objects and events, but vague impressions are of little use to science. In order to form precise conclusions, these relationships must be described objectively, but this demands that we describe the process of measuring. We must learn to distinguish between those things that are physically significant, and those that are relative to how they are measured.

The most obvious datum is position. Any object in the universe has a location in space at each instant of time; these points in space and time are labeled with **coordinates**. The customary notation for coordinate locations is (x, y, z, t) , where x , y , and z represent the spatial quantities, and t represents time. (If we simplify matters by working with only one spatial dimension, we shall refer to its coordinate as x .) Coordinates are merely convenient labels, not physical attributes of space or time, so the symbols and units chosen are arbitrary. The coordinates of a point have no intrinsic significance; their only importance lies in the relationships between two sets of coordinate values, such as relative locations. We measure space by means of a standard, which we shall generically call a ruler, regardless of what it might actually be. We measure time by means of clocks, where a clock could be any standard periodic physical process and need not literally refer to a wristwatch. Distance and time intervals have physical significance, but whether we measure a distance interval in inches, yards, or meters is not important. A measurement in meters is merely an expression relative to an arbitrary standard, but the distance itself represents a real, reproducible, quantifiable measurement.

Coordinate system defined

Since measurements of separations in space and time are among the most important, we will concentrate initially on understanding the meaning of these quantities. As a specific example, if we wish to know the

The Pythagorean rule computes distance from coordinate separations

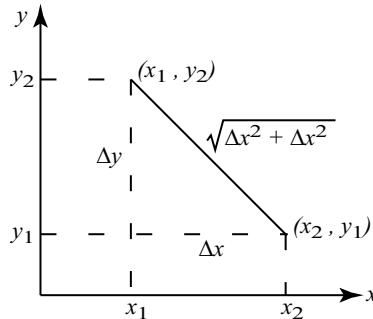


Fig. 6.5 The Pythagorean rule. The distance between points (x_1, y_1) and (x_2, y_2) is given by the square root of the sum of the squares of the coordinate separations, Δx and Δy .

velocity

acceleration

distance between two points in space, we may begin by laying down coordinate lines that run at right angles to one another, constructing a grid in x and y . Next we assign spatial coordinate locations to each point, say (x_1, y_1) for point 1, and (x_2, y_2) for point 2; then we determine the difference³ between those coordinates, $\Delta x = x_2 - x_1$ and $\Delta y = y_2 - y_1$. The Pythagorean theorem enables us to find the desired distance by summing the squares of the two sides to obtain the square of the distance between points 1 and 2; specifically:

$$s^2 = \Delta x^2 + \Delta y^2, \quad (6.1)$$

where we have used the symbol s to indicate distance. The Pythagorean theorem has an obvious generalization to three dimensions—the diagonal of a cube rather than of a square—but for the present illustration, the familiar two-dimensional version is sufficient.

Quite often we want to know the length of time required to travel a specified distance. The quantity that describes the change of position with time is **velocity**. Velocity is a vector, meaning that it has a direction as well as a magnitude associated with it; the magnitude of the velocity is the **speed**. For example, if you travel 20 kilometers in half an hour, you can say that your speed is 40 kilometers per hour. If you further state that you traveled from east to west, then your velocity is specified as 40 kilometers per hour toward the west. If you drive on a winding road at a constant speed of 40 kilometers per hour, your velocity changes with each turn of the wheel. The distinction between speed and velocity can be very important, and should be kept firmly in mind. In general, velocity is defined as a *derivative*, that is, the rate of change, of the three-dimensional position vector. However, by confining our attention to motion along one specific direction (x) we can write this as $v = \Delta x / \Delta t$, simplifying the mathematics without much loss of qualitative content. Velocity is the change in space position, divided by the accompanying change in time.

A change of velocity with time is an **acceleration**, and is written $\Delta v / \Delta t$. Since acceleration is defined in terms of velocity, it too carries directional information. An object may have both an acceleration and a velocity, and they need not be in the same direction at all. If you jump from the ground, your velocity is initially in the upward direction; but the gravitational acceleration is directed down toward the Earth, which is why you eventually reverse your motion and fall back to the surface. Orbits provide another example. The velocity of an orbiting planet is nearly perpendicular to the line between the planet and the Sun, whereas the gravitational acceleration is directed from the planet toward the Sun. In the case of purely circular motion of any kind, the velocity and the acceleration are exactly perpendicular to one another. Moreover, since acceleration is the change in velocity, there may exist an acceleration

³The symbol Δ is the standard mathematical notation for the concept of change in a quantity. Thus the expression Δx indicates “the change in the spatial coordinate x ”; the Δ and the x are inseparable in this context.

even if the speed never changes. Riding along that winding road at a very steady speed of 40 kilometers per hour, you will nevertheless feel an acceleration at each curve.

It may seem that velocity and acceleration are very similar quantities, both being descriptions of how something changes with respect to time, but there are important physical distinctions between the two. Newton's second law, $F = ma$, tells us that a force is required to produce an acceleration. By Newton's first law, in the absence of a force a body in a state of uniform motion will continue in that same state indefinitely. Stated simply, uniform motion means constant velocity. (Rest is a special case for which the velocity is zero.) Uniform motion is the natural state and will last indefinitely; only a force can cause an acceleration. This means that there is some attribute of a body, its **inertia**, which causes it to resist changes in its velocity. How do we quantify inertia? We do so through a property we call **mass**. Inertial mass is *defined* by Newton's second law; we measure mass by applying a known force and observing the resulting acceleration. We refer to unaccelerated motion as **inertial motion**. Thus any uniform motion is an inertial motion.

There is a real, physical difference between inertial motion and accelerated motion. But if our units are arbitrary, how may we determine whether a motion is accelerated or inertial? Suppose we had a measuring device that changed its length-scale, or a clock whose mechanism made it run at different rates. With this ruler and clock, it would seem that velocity is continually changing. How could we distinguish a measurement made with such odd measuring devices from true accelerated motion? There are two ways to answer this. First, we can note that mathematically, the properties of true accelerated motion are expressed by Newton's second law, and this relationship does not depend upon our measuring units. As a more trivial example, if someone tells you to time your heartbeat but hands you a defective watch, you should not conclude that your heart is malfunctioning. Second, we can appeal to experience: the difference between inertial motion and accelerated motion is *palpable*. Acceleration requires a force, and this has consequences for objects such as the human body. Sometimes acceleration is measured in units of g , the gravitational acceleration at the surface of the Earth. Therefore, accelerations are often called, loosely, g -forces. Pilots of fighter planes must wear special g -suits because they may experience very high accelerations in a tight turn or dive. (Circular motion, even in the case that only part of the circle is traversed or the radius of the circle is changing, means that there is an acceleration and hence a force.) Such large forces can cause humans to pass out, or can even be fatal if strong enough.

To clarify the distinction between acceleration and distorted units of measure, we must introduce the concept of a **frame of reference**. The frame of reference is a system of coordinates attached to an observer whose viewpoint we are considering. Suppose you define the origin of an (x, y, z) coordinate system coincident with your navel, and time coincident with the watch on your wrist. These specifications define a frame

Inertial motion is unaccelerated motion

Newton's first and second laws hold in an inertial reference frame

The Rotor: an example of an accelerated frame

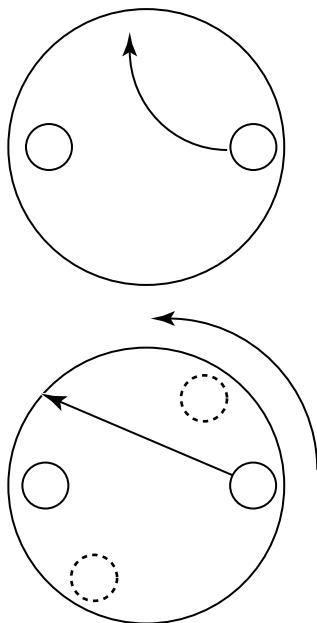


Fig. 6.6 (top) As seen in the Rotor reference frame, the ball appears to curve as if it were accelerated. (bottom) In the frame of an external observer, the ball moves in a straight line, consistent with unaccelerated motion, while the riders rotate from their original locations (solid circles) to new locations (dashed circles).

of reference. With respect to the coordinates attached to your body, you are always at rest, since your coordinates move with you; thus this frame is defined to be your *rest frame*. How is your rest frame related to other frames, such as the frame defined by the distant fixed stars? In particular, what is your state of motion relative to the fixed stars? Are you at rest, moving with a constant velocity, or undergoing acceleration? This has implications for your frame of reference. As a specific example, suppose you and your coordinate system are in deep space; you are unaccelerated. What does this imply? Your accelerometer reads zero, and you feel no forces acting on your body. You float along at a constant velocity, in inertial motion. In such a case, you reside within a very special reference frame, called an **inertial reference frame**. An inertial reference frame is *any frame in which a free particle executes uniform motion*, that is, moves at a constant velocity, as specified by Newton's first law. In an inertial frame, a particle set into motion at constant velocity would continue in such uniform motion indefinitely.

A noninertial reference frame is thus any frame that is not inertial, but what are the physical implications of such a frame? Most obviously, this means that forces are acting upon all objects within the frame; but how do forces affect the reference frame itself? An example might clarify these issues. Many amusement parks and fairs feature a ride known as the "Rotor" or some similar name. The Rotor consists of a tube whose inner walls are covered with a rough material, such as burlap. The riders stand against the walls of the cylinder, and it begins to rotate. When it reaches a certain angular speed, the floor drops about a meter, and the riders adhere to the walls. It is the friction between clothing and the burlap that prevents the riders from sliding down, but the force that presses them against the burlap comes from the acceleration they experience due to the circular motion. What happens to the motion of free particles within such a reference frame? Suppose that while you are riding the Rotor, you decide to play catch with a friend directly across from you; you toss a ball toward your friend. What happens? The ball does not travel toward your friend, but veers off to the side. But suppose another friend watches from above, outside the Rotor, as you throw the ball. Your overhead friend insists later that the ball, once released from your hand, flew in a straight line, as Newton's first law predicts. In this example, in *your* frame of reference, the ball curved. In your friend's frame, the ball traveled in a straight line.

The rotating frame within the Rotor is, clearly, *not* an inertial reference frame; the ball, traveling with constant horizontal velocity once it left your hand, appeared to curve with respect to your frame of reference. Moreover, you feel forces; you are pressed against the walls of the cylinder, and if you carried an accelerometer with you, it would show a nonzero reading. The Earth, like the Rotor, is rotating, and hence it is not an inertial reference frame, although for many purposes, such as measurements on the scale of a laboratory, it is approximately so, since its rotation is relatively slow. But what about physical phenomena for which the rotation creates a significant effect? Is there any way to make



Fig. 6.7 Coriolis forces affect the motions of storms upon Earth. In this photograph, taken by the astronauts of *Apollo 17*, Africa and the Middle East are visible, with the island of Madagascar near the center of the view. The curving motions of the winds are made visible by clouds; a particularly well developed storm system can be seen at the upper right, off the horn of Africa. (NASA.)

sense of measurements within a noninertial frame? If we insist upon using a rotating frame of reference such as the Earth, we may write Newton's second law with various modifications that take the rotation into account; the resulting equation brings in two so-called *fictitious forces*, the *centrifugal force* and the *Coriolis force*. Forces such as these two are called fictitious because they exist only due to our use of the wrong, that is, a noninertial, frame of reference; they are also called **inertial forces** because they arise from the acceleration of the frame of reference. They are perfectly valid within the accelerated frame, however, and it is often easiest to use such noninertial frames. For example, the inertial forces appear in the equations used in meteorology and oceanography; no one in those disciplines would even think of trying to use a reference frame fixed to the distant stars.

Centrifugal force is directed away from the center of rotation. In the Rotor example, the riders feel a centrifugal force pushing them against the wall. The Coriolis force causes a moving object to curve in the rotating frame; it causes the ball to be deflected from a straight line in the Rotor frame. On the Earth it causes storms to form cyclones, which are clearly evident by their cloud patterns in satellite photographs. Air parcels moving toward a center of low pressure are deflected by the Coriolis force, causing them to swirl around the low; in the Northern Hemisphere this induced rotation is counterclockwise, while in the Southern Hemisphere it is clockwise.⁴ It is possible to detect the Earth's rotation

Inertial forces arise from an accelerated reference frame

⁴ Although the Coriolis force affects all motion on the Earth, it is, despite many urban legends, nearly impossible to observe in an ordinary bathtub or bucket. On those scales the force is much too small to see, for practical purposes. Water draining from a tub swirls mainly because of effects that are far more significant than the Coriolis force at that scale; it would be as likely for water to drain from a kitchen

directly by measuring the Coriolis force. The magnitude of the Coriolis force is proportional to the rotation rate of the frame of reference, multiplied by the speed of the moving object. The rotation rate of the Earth is small, so that Coriolis forces are also small, and are not easily observable on everyday length-scales. The Foucault pendulum, often a prominent fixture in science museums, is one exception. This pendulum consists of a heavy bob suspended from a bearing that allows it to turn in any direction. Within the Earth-bound (noninertial) reference frame, the pendulum bob is deflected by the Coriolis force, so that it appears to rotate; the precise rate of rotation depends upon the latitude on the Earth at which the bob is located. A visit to a well-equipped science museum can provide a first-hand demonstration of the Earth's rotation.

Gravity is another effect that prevents our frame of reference on the surface of the Earth from being inertial. We are so accustomed to living in a gravitational field that we may have difficulty in visualizing a true inertial frame, but from our definitions it is clear that the presence of gravity creates a noninertial frame. Any object moving near the Earth will be affected by a gravitational acceleration. Like the centrifugal and Coriolis forces, we are aware of gravity and can account for it when we write our equations of motion for objects moving near the surface of the Earth. Gravity also acts only vertically, so that horizontal motions and forces are unaffected by it. (In fact, the direction of gravity *defines* the vertical.) If a reference frame is moving at constant velocity in a gravitational field, then motions occur within that frame of reference *exactly* as they would at rest, with the effects of gravity included. An airplane can approximate such a frame of reference if the air is smooth. Suppose you are riding in a jet at cruising speed and altitude, with no atmospheric turbulence in your path. The flight attendant hands you a can of soft drink and a cup; you pour the soda into the cup exactly as you would if you were sitting at rest in your kitchen at home. The cup and the stream of liquid share the same constant horizontal velocity, so no effect of that velocity can be detected within your frame. If the airplane accelerates, however, either by speeding up or by changing its direction, you are likely to spill the soda as you attempt to pour it. This is similar to the arguments used by Galileo and others to demonstrate that motion is not always detectable from within the moving frame of reference.

Since gravity is ubiquitous throughout the universe, what, then, would constitute a truly inertial reference frame? One example would be a spaceship traveling at constant velocity in deep space, where gravity is, locally, negligibly small. Another example of an inertial frame is one that

Our everyday reference frame is non-inertial due to gravity

Freefall as inertial motion

sink in a clockwise sense in the Northern Hemisphere, as in a counterclockwise sense. However, the Coriolis force *does* significantly affect the trajectories of long-distance artillery shells, as some British naval gunners found to their embarrassment in a conflict in the Southern Hemisphere; they used Northern Hemisphere tables of the Coriolis force for their targeting corrections, but this force acts in the opposite direction in the Southern Hemisphere. Long-range shells fired with the incorrect aiming missed their targets by miles.

is *freely falling* in a gravitational field. What is special about freefall in a gravitational field? Recall that mass appears both in Newton's second law and in Newton's law of universal gravitation. We have even written these two masses with the same symbol m , but they are really two distinct concepts. In the second law, mass is a measure of the inertia, or the resistance to acceleration. In the law of gravity, mass is a measure of gravitational charge, analogous to the role of electric charge in the theory of electromagnetism. If the gravitational mass is the same as the inertial mass, then we may combine these two equations and cancel the mass of the test object. For any object falling in the gravitational field of the Earth, we may write

$$g = \frac{GM_E}{R_E^2}, \quad (6.2)$$

which does *not* depend on the mass of the body. This is the mathematical expression of the experimental result that all objects fall at the same rate in a gravitational field. In the absence of any nongravitational force such as air resistance, a feather and a cannonball dropped from the same height at the same time will hit the ground together. Apollo astronaut David Scott performed exactly such an experiment on the airless Moon during the Apollo 15 mission in 1971, dropping a feather and a hammer at the same instant. Both fell with the same acceleration and struck the surface simultaneously.

The independence of gravitational acceleration from the inertial mass provides the solution to a famous trick question of physics. A hunter aims at a monkey who is holding the branch of a tree. Just as the hunter fires, the monkey lets go of the branch in an attempt to evade the bullet. Does the bullet hit him? The answer is yes, if the hunter's aim is accurate, because both the bullet and the monkey fall in the vertical direction at exactly the same rate, despite their large difference in mass. Because in freefall everything falls together at exactly the same rate, gravity is effectively canceled out; all motions relative to the freefalling frame will be consistent with Newton's laws. This seemingly innocent and obvious equivalence of inertial and gravitational mass will be shown to have some amazing and profound consequences through the general theory of relativity.

The relativity of space and time

Given that inertial frames of reference exist, why are such frames important? When making measurements using a coordinate system, it must be possible to distinguish those things that are physically significant from those that are related only to the specific reference system by which they are measured. We have suggested such a distinction by emphasizing that acceleration has physical consequences, independent of the coordinate system used to measure it. We can clarify matters even further by means of some definitions, beginning with an **inertial observer**. An inertial observer is simply an observer whose rest frame

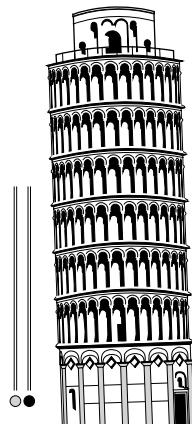


Fig. 6.8 Two balls of different masses accelerate at the same rate when dropped.

Invariance and relativity: agreement and disagreement in measurement

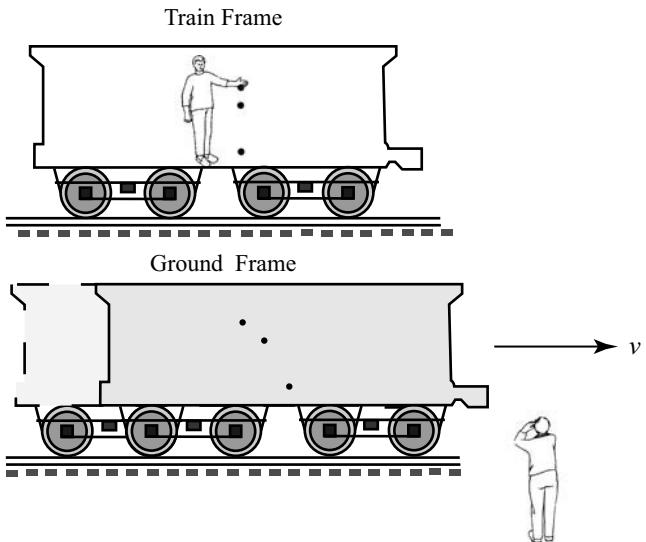


Fig. 6.9 Path of a dropped ball, as seen (top) in the frame of a train moving at constant velocity and (bottom) in the frame at rest with respect to the Earth. *Relativity* describes how to relate measurements made in one inertial frame to those made in another.

is inertial. Next comes the concept of **invariance**. A quantity is said to be invariant if all inertial observers would obtain the same result from a measurement of this quantity. On the other hand, a quantity is said to be *relative* if different inertial observers obtain different results from their measurements. **Relativity**, which is a general term and does not apply only to Einstein's theory, tells us how to relate observations made in one inertial frame of reference to observations in another such frame.

As a first example, let us consider the frame of a train moving at constant velocity. One of the passengers drops a ball onto the aisle. Another passenger who observes the fall of the ball will see exactly what she would see if the ball were dropped on the surface of the Earth; the ball lands at the feet of the person who dropped it. Suppose that another observer, who is at rest with respect to the Earth, watches the same ball as the train goes by. The Earth-based observer measures the path of the ball, relative to his own frame, to be a parabola, since the ball shares the horizontal velocity of the train. Both observers agree that the ball accelerated downward due to the force of gravity. Both agree on the magnitude of that force, on the mass of the ball, on the value of the acceleration, and on the length of time required for the ball to fall. Both observers can apply Newton's laws of motion to compute the theoretical path of the ball. However, they disagree on the velocity of the ball, the path it took while falling, and its final position. These differences are all attributable to the motion of the train. The quantities acceleration, mass, force, and time interval are invariant. The observers disagree on the coordinates, because they are using different coordinate frames; they also obtain different results for the position and the net velocity of the ball at any given time. Quantities such as coordinates, position, and velocity are relative.

Inherent in this example is the assumption that space and time are absolute. All observers agree on space and time intervals; that is, one second of time and one meter of distance are the same in all inertial frames. Moreover, all inertial frames are equivalent. There is no absolute motion per se, no one correct inertial frame that is better than any other. Since all inertial frames are equally valid, we need only find the procedure for relating measurements in one frame to the measurements in another, thereby accounting for the relative quantities. The equations that relate measurements made in one Newtonian inertial frame to those made in another are called, collectively, the *Galilean transformation*. They are very simple and intuitive; basically, the equations of the Galilean transformation simply adjust the observed velocities by the relative velocity between the two frames. In our example above, if the train is moving with speed v_{train} toward the west, as measured by the observer on the ground, and the passenger throws the ball down the aisle toward the west with horizontal speed v_{ball} , as measured by an observer on the train, then the horizontal velocity of the ball, as measured by the observer on the ground, is

$$v_{\text{ground}} = v_{\text{train}} + v_{\text{ball}} \quad (6.3)$$

toward the west. On the other hand, if the thrower faces the back of the train and tosses the ball toward the east with horizontal speed v_{ball} , then the horizontal velocity of the ball measured by the ground-based observer is

$$v_{\text{ground}} = v_{\text{train}} - v_{\text{ball}} \quad (6.4)$$

toward the west.

Since physical laws are intended to describe some objective properties of the universe, we can see that they must be invariant under the transformation from one inertial observer to another; otherwise they would depend upon the coordinate system used to make measurements, and coordinate systems, as we have emphasized, are purely arbitrary.

Galilean relativity is the formal statement that Newton's laws of motion are invariant under the Galilean transformation. That is, Newton's laws work equally well, and in the same manner, in all inertial reference frames, if those frames are related by the Galilean transformation. Thus, if Galilean relativity gives the correct relationship between inertial frames, Newton's laws provide an accurate description of the laws of mechanics, since they do not change their form under a Galilean transformation. Note too that since Newton's laws operate precisely the same in all inertial reference frames, no experiment can distinguish one such frame from another; this implies that you can never tell if you are "really moving" or "really at rest," as long as your motion is unaccelerated. Perhaps there could be some absolute cosmic frame of rest, although there is nothing in the Newtonian universe to suggest such a thing, and the introduction of any special frame of reference would tend to vitiate the spirit of relativity.

All inertial reference frames are equally valid

Measurements can be transformed from one frame to another

Galilean relativity

A fly in the ointment

The theory of electromagnetism

The ether hypothesis

During the middle part of the 19th century, scientists were fairly certain that all of physics must be invariant under the Galilean transformation. However, the laws of physics were still being uncovered; one area of especially active research was electricity and magnetism. In the 1860s, James Clerk Maxwell developed a theory of electricity and magnetism that showed that these two forces were actually manifestations of one electromagnetic force. A consequence of Maxwell's equations was that fluctuating, time-varying electromagnetic fields traveled through space at the speed of light. It soon became clear that this electromagnetic radiation *was* light itself. Maxwell's equations, which describe the evolution of electric and magnetic fields, depend specifically upon a speed: the speed of light. Yet speed is a quantity that is relative under Galilean transformations, so Maxwell's equations are also relative under these transformations. When the equations were developed, however, this lack of Galilean invariance was not immediately troubling to most physicists. Scientists of the time understood waves in matter, such as elastic waves or sound waves. All such waves require a medium in which to propagate, and the speeds that describe these waves, such as the speed of sound, are specified with respect to the medium through which the wave travels. The net velocity of the wave, as seen by an observer not moving with the medium, is the vector sum of the velocity of propagation plus the velocity, if nonzero, of the medium. Since all the waves familiar in the middle 19th century were of this nature, the reasoning of the day concluded that light too traveled through a medium, which was called the **luminiferous ether**, or just the *ether*. This ether has nothing to do with the volatile chemical substance of the same name, nor is it the same as the celestial ether of Aristotle; the luminiferous ether had no other reason for its existence than to provide the expected medium for the propagation of light.⁵ It had no particular tangible properties of its own; it was massless and invisible. This seems rather peculiar; why should the universe contain this strange substance with such a specialized function? After all, air does not exist solely to carry sound waves. But so strong was the mechanical picture of waves in the minds of 19th-century scientists that no other alternative was seriously entertained.

It was thus assumed that Maxwell's equations were valid only in the frame of the ether. Many physicists of the time even concluded that the rest frame of the ether could be identified with the Newtonian absolute space. But if the ether has a frame, then that must be some kind of *preferred* frame of reference, which presumably fills all space. As such, it must be possible to detect the ether through its special frame of reference; in particular, a carefully designed experiment should be able to measure the motion of the Earth through the luminiferous ether. Once the ether was observed, it was thought, the theory of electromagnetism

⁵A similar ether had been proposed earlier to account for the transmission of gravity over distance.

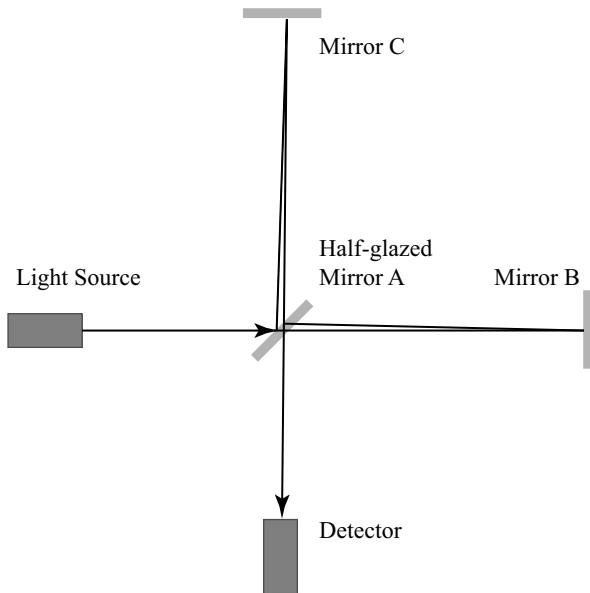


Fig. 6.10 Schematic illustration of a Michelson–Morley interferometer experiment. Light from a source travels to a half-glazed mirror (A) that splits the light beam into two, sending light down both arms of the apparatus to mirrors (B) and (C). If the round-trip time along AB differed from that of AC due to differences in the speed of light with direction, the observer would see interference fringes when the beams recombined.

would be complete; together with Newton's laws, the description of the fundamental properties of nature would also then be finished. In 1887 Albert Michelson and Edward W. Morley set out to measure the motion of the Earth with respect to this frame. Michelson and Morley set up an experiment in which a beam splitter broke a beam of light into two. One half of the original beam was sent in one direction, struck a mirror, and was reflected back to another, angled mirror. The other half of the beam traveled precisely the same distance perpendicular to the first direction, where it was also reflected and returned. The experiment asked whether the transit time was equal for the two perpendicular round trips. If light behaved like a mechanical wave, the experimental setup would be analogous to two swimmers in a river, one traveling across the current and back, and the other swimming the same distance downstream and then returning upstream. The swimmer who had only to cross the current twice would complete the trip faster than the swimmer who had to battle the current on the way back. The difference in swimming time could be used to derive the flow speed of the river.

Michelson and Morley measured the transit time for the light by recombining the light beams upon their arrival, thus superposing the two light waves. If the light beams had different round trip times, they would be out of phase when recombined. Adding light waves with different phases results in alternating constructive and destructive interference, producing a pattern of light and dark known as *interference fringes*.⁶ A device to observe such fringes is called an **interferometer**; this particular experimental setup is known as a Michelson–Morley interferometer. The apparatus was constructed so that it could be rotated, turning one

Measuring motion with respect to the ether

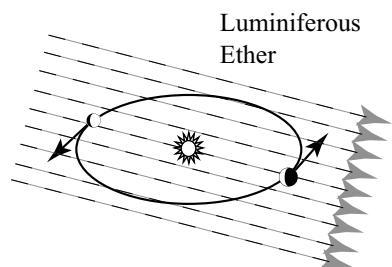


Fig. 6.11 At different times of the year the Earth would move in different directions with respect to the ether.

⁶Wave properties such as interference are described in Chapter 4.

The “failure” of the Michelson–Morley experiment

arm and then the other toward the direction of motion of the Earth. Since the speed of light plus ether was expected to differ for the two arms in a predictable way as the device was rotated, the change in the interference fringes would provide the difference in the travel time of the light along the two paths, and hence the velocity of the Earth with respect to the ether. That, at least, was the idea. To their great surprise, however, *no* difference in light travel times was observed. Michelson and Morley repeated their experiment numerous times and at different times of the year. In the end, they determined that the velocity of light was the same, to less than 5 kilometers per second, in the two mutually perpendicular directions. This result was well within their experimental error; therefore, the outcome of their experiment was the declaration that the speed of light was equal in both directions. While this might seem at first glance to be an experimental failure, their null result was one of the most important experimental observations of the late 19th century.

By the last quarter of the 19th century, the universe had expanded dramatically, both in space and in time. The distance from the Earth to the Sun had been accurately measured about a century before, in 1769, finally setting the absolute scale of the solar system; the solar system alone was discovered to be much larger than the size of the entire universe in the Aristotelian cosmos. Evidence was mounting that the Earth was several billions of years old. Change and evolution throughout the universe were becoming an accepted paradigm. Newtonian mechanics seemed to give humanity a glimpse of the architecture of the universe itself. Physicists felt they had every reason to feel proud, perhaps even a little smug. And yet, a few pieces of the electromagnetic theory still could not be made to fit. The Michelson–Morley experiment left physicists in some disarray for nearly twenty years. Except for a few small difficulties such as the minor confusion about the ether, physics had seemed to be more or less wrapped up.⁷ Yet this seemingly small inconsistency led directly to the development of a new and startling theory, and a new way of looking at space and time that will form the foundation for our modern cosmological theories. We turn now to Einstein’s theory of relativity.

Einstein

Albert Einstein was born in Ulm, Germany, the son of a less-than-successful businessman. An unspectacular, though not untalented, student, he left Germany in his teens and traveled through Italy. Eventually he settled in Switzerland, where he attended the Swiss Federal Institute of Technology, finally obtaining his doctorate in 1900. Unable to find employment as a scientist, he accepted a position as a patent examiner with the Swiss Patent Office in Bern. In his later life he reminisced

⁷Another “minor” problem was explaining the blackbody spectrum. The resolution of that problem led to the creation of quantum mechanics.

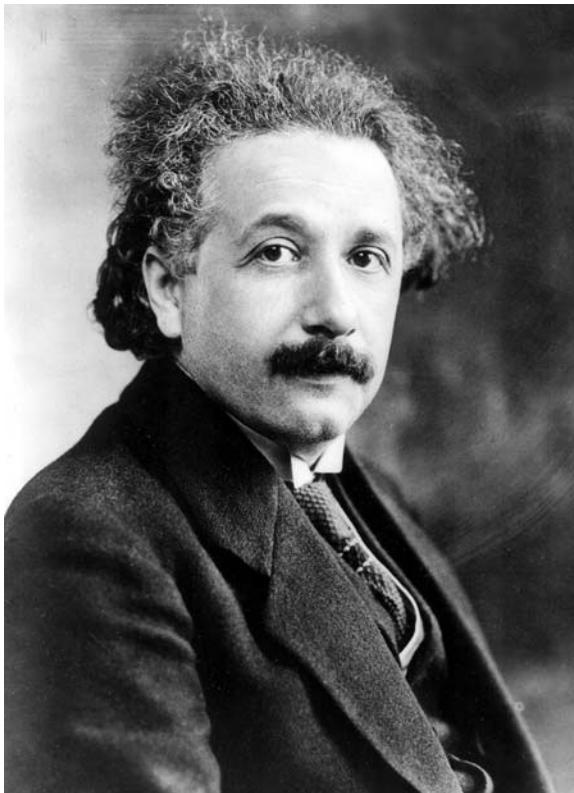


Fig. 6.12 Albert Einstein (1879–1955). Best known for formulating the special and general theories of relativity, Einstein made many fundamental contributions to the development of quantum mechanics as well. (Courtesy of Yerkes Observatory.)

nostalgically about his days as a patent clerk. He enjoyed the work of evaluating patent applications, and his life as a scientific outsider seems, if anything, to have stimulated his creativity. In 1905 he published three epochal papers. One was a work on Brownian motion, the jiggling of tiny particles due to the many impacts of molecules of air or water upon them. Another was his explanation of the photoelectric effect, a then-mysterious phenomenon that occurs when light strikes the surface of a metal. This paper employed and elaborated upon the quantum theory of radiation developed a few years previously by Max Planck to explain the blackbody spectrum. Einstein's grand hypothesis was that light itself was quantized; we now refer to a quantum of light as a *photon*. The explanation of the photoelectric effect was one of the earliest applications of quantum mechanics, and eventually won Einstein the Nobel Prize in physics. The third paper, *Zur Elektrodynamik bewegter Körper* (On the electrodynamics of moving bodies), published in the German scientific journal *Annalen der Physik*, laid out the special theory of relativity.

The special theory of relativity wrought dramatic changes in our view of the universe. No longer could we imagine the absolute, pristine space and time of Newton. Space and time were not the stage upon which the drama of dynamics unfolded; they became actors in the play. The special theory showed that the electromagnetics of Clerk Maxwell was, in fact,

The life of Einstein

The special theory of relativity

more accurate than the mechanics of Newton. It does not denigrate Newton's great achievements in the least to discover that his physics was not quite right; he could not have arrived at the correct formulation even with his intimidating genius, as the necessary understanding of electromagnetics was lacking in his day. Newton's mechanics is an approximation, valid only in the limit of speeds that are very small relative to the speed of light. Since essentially all Earthly motions occur at such speeds, certainly for all macroscopic objects, Newton's theory seemed completely adequate. The need for the special theory of relativity was not perceived until a contradiction was discovered with what seemed, at first glance, to be a completely separate arena of physics.

Einstein's search for a general theory of relativity

After his triumph with mechanics, Einstein turned to gravitation. This proved a tougher nut to crack, and occupied Einstein for the next ten years. By then, he had become a member of the scientific establishment, securing prestigious positions at universities in Prague, Zurich, and finally Berlin. Although he arrived quickly at the physical foundations of what became the general theory of relativity, the mathematical representation of the ideas was far from obvious, and Einstein reached many dead ends. Finally, around the time of the First World War, his friend Marcel Grossman introduced him to a branch of mathematics known as Riemannian geometry. Einstein found his answer there; the equations of general relativity were published late in 1916. Almost immediately, they were applied to cosmology, first by Einstein himself, in 1917, and later by scientists such as Alexander Friedmann, Willem de Sitter, and Georges Lemaître.

It is unfortunate that both special and general relativity have acquired such an intimidating reputation. The special theory requires no more than algebra for a basic understanding of its workings, although details of its application demand somewhat higher mathematics. The general theory is, of course, more complex, and cannot be fully understood without higher mathematics; the fundamental ideas, however, are not intrinsically difficult. The real impediment to the understanding of both theories is not the mathematics, but the new way of thinking they demand. Our intuitions often mislead us in our attempts to understand even Newtonian mechanics. The theories of relativity require a mental flexibility that the complacent of mind may not be willing to attempt. Yet a little effort can provide a basic understanding of these great ideas that so significantly shaped physics in the 20th century.

Chapter Summary

A model of the universe deals with the physical universe and its contents. In modern physics and cosmology, space and time are themselves physical and are part of the universe; the big bang did not occur in a pre-existing space

and time. Because science must deal with physical entities, the issue of the creation of the universe is necessarily metaphysical. Our existence, and the special properties that the universe must possess in order to permit this ex-

istence, is an intriguing mystery. The anthropic principle has been used to argue that the presence of life constrains the universe or determines why the universe is as it is, but at the present this remains only speculation.

Modern cosmological models are founded on the cosmological principle, which asserts that the universe is *homogeneous*, that is, has the same average properties everywhere, and *isotropic*, meaning there are no special directions in space. We observe that on the largest scales the universe appears the same in all directions; hence it is isotropic. Unless we are at the center of the universe, it follows that the universe must also be homogeneous. The universe has neither a center nor an edge.

Observations require a precise system of coordinates and units to standardize measurements. A given set of coordinates is a reference frame, and a reference frame whose origin is unaccelerated is an inertial reference frame. Unforced velocities measured in an inertial frame obey Newton's laws. Measured in an accelerated frame,

such motions appear to be accelerated; we say that they are influenced by frame-dependent inertial forces, such as the centrifugal or Coriolis forces. Gravity acts like an inertial force, as the acceleration produced by gravity on an object is independent of its mass.

Galileo realized that if everything on the Earth shared the same overall motion, then that motion would be undetectable. This leads to the conclusion that there is no absolute frame of rest, a condition that is required for Newton's first law of motion. Galilean relativity states that one inertial frame is completely equivalent to another; reality should not depend on the arbitrary frame in which it is studied. Newton's laws must be the same in any inertial frame. However, Maxwell's theory of electromagnetism could not be made to conform to Galilean relativity. This inconsistency led Einstein to a new version of relativity that maintains the underlying concept, the complete equivalence of all inertial frames, but replaces Newton's laws with more general laws of motion.

Key Term Definitions

anthropic principle The observation that, since we exist, the conditions of the universe must be such as to permit life to exist.

Copernican principle The principle that the Earth is not the center of the universe.

isotropy The property of sameness in all directions, as in an isotropic geometry.

homogeneity The property of a geometry that all points are equivalent.

cosmological principle The principle that there is no center to the universe, that is, that the universe is isotropic on the largest scales, from which it follows that it is also homogeneous.

perfect cosmological principle The principle that the universe is unchanging, that it is homogeneous in time as well as in space. Refuted by the direct observation that the oldest objects in the universe are not like those in our immediate surroundings.

coordinates Quantities that provide references for locations in space and time.

velocity The rate of change of displacement with time. Velocity includes both the speed of motion and the direction of motion.

speed The magnitude of velocity.

acceleration The rate of change of velocity with time.

inertia That property of an object which resists changes in its state of motion.

mass That property of an object which causes it to resist changes in its state of motion; also, that property which generates gravitational attraction.

inertial motion Motion free of any force, that is, motion at constant velocity.

frame of reference The coordinate system to which a particular observer refers measurements.

inertial reference frame A reference frame in which a free particle experiences no force.

inertial force A force arising from the acceleration of an observer's frame of reference.

inertial observer An observer occupying an inertial frame of reference.

invariance The property of remaining unchanged under a transformation of the frame of reference or the coordinate system.

relativity The rules relating observations in one inertial frame of reference to the observations of the same phenomenon in another inertial frame of reference. Casually applied only to the Einsteinian

special theory of relativity, but actually a more general term.

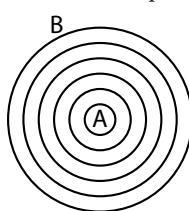
Galilean relativity The transformation from one inertial frame of reference to another in the limit of very small velocities and very weak gravitational fields.

luminiferous ether A supposed medium for the transmission of light. The concept was rendered superfluous by the special theory of relativity early in the 20th century.

interferometer A device that carries out some measurement by detecting wave interference.

Review Questions

- (6.1) Describe the weak and strong anthropic principles. What philosophical assertion does each make about the universe? What do you think about them?
- (6.2) Why is it not a scientifically valid question to ask what happened before the universe came into existence?
- (6.3) Flatlanders live in a two-dimensional universe. Suppose such a universe were described by the figure, such that all matter is confined to the indicated rings. Does this universe appear isotropic to an observer at point A? To an observer at B? Is this universe homogeneous for either observer? What would the observer at A conclude from applying the Copernican principle? Explain your answer. Draw an example of a universe that is homogeneous but nowhere isotropic.
- (6.4) Is the cosmological principle consistent with the existence of a center or an edge to the universe? Explain.
- (6.5) Explain the distinction between the cosmological principle and the perfect cosmological principle.
- (6.6) We have mentioned that galaxies are grouped into clusters. How can the existence of such clusters be consistent with a homogeneous universe?
- (6.7) Explain the distinction between invariant and relative quantities.
- (6.8) An airplane is traveling at 300 mph toward the west. A rambunctious child seated in front of you throws a ball toward the tail of the aircraft, that is, toward the east, at 6 mph. According to Galilean relativity, what is the speed of the ball relative to an observer in the airplane? Relative to an observer at rest on the surface of the Earth?
- (6.9) You wake to find yourself in an airplane with all its windows covered. Is there any experiment you can perform to determine whether you are flying with a uniform velocity, or at rest on the runway? (Ignore external effects such as engine noises, which could be simulated as a diabolical plot to trick you.) If the airplane changed its velocity, could an experiment show this? If so, give an example of an experiment you might perform that could detect an acceleration of the airplane.
- (6.10) Why did the appearance of the speed of light in Maxwell's equations create a problem for Galilean relativity theory?



The Special Theory of Relativity

7

Key Terms:

The grand aim of all science is to cover the greatest number of empirical facts by logical deduction from the smallest number of hypotheses or axioms.

Albert Einstein

Einstein's relativity

In some faraway galaxy, an advanced civilization has mastered space travel. The crew of one of their starships discovers an asteroid on a collision course with one of their world's space outposts, which would surely be destroyed if the asteroid collides with it. The asteroid is at a distance of 3,000,000 km, as measured by the sensors on the starship. The starship is flying toward the asteroid at nearly the speed of light as seen from the space station. The ship fires its laser cannon at the asteroid. What would the captain of the ship observe? What would the officer on duty at the station see? When will the laser light beam reach its target?

If light were analogous to sound waves, we could use Galilean relativity to find the correct answers. Sound waves are waves of pressure moving through a fluid, such as the air. Because they are waves in a medium, they move at a specific velocity (the speed of sound) relative to the medium. Wind carries sound along with it, and the total speed of the sound relative to the ground is the speed of the waves relative to the air, plus the speed of the wind, taking directions of motion into account. If we regard light as moving through some medium, which was historically called the luminiferous ether, then the light waves will always move at the speed of light, *relative to the ether*. We are now in a position to deduce what will happen in the spaceship problem posed above; we need only know how fast the spaceship is moving with respect to the ether, just as we might compute how rapidly sound waves would travel if emitted by a speaker mounted atop a moving vehicle.

In Newtonian cosmology, space and time are absolute, and the same for all observers. If the ship and the laser beam are traveling in the same direction, Galilean relativity tells us that we should simply add the speed

- Lorentz transformation
- length contraction
- relativity principle
- event
- simultaneity
- time dilation
- proper time
- boost factor
- principle of reciprocity
- proper length
- rest energy
- space-time
- Minkowskian space-time
- space-time diagram
- worldline
- space-time interval
- timelike
- spacelike
- lightlike
- lightcone
- past
- future
- elsewhere
- principle of causality

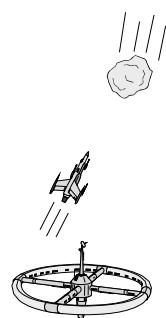


Fig. 7.1 An asteroid on a collision course with a space station.

of the ship to the speed of light, which is about $300,000 \text{ km s}^{-1}$ in the vacuum, to obtain the net speed of the laser cannon beam. Assuming that the ship is traveling at 99.99% of the speed of light relative to the ether, we would compute that the laser beam will hit the asteroid 5.0025 seconds after firing. Is this the correct answer? What would happen if the asteroid were behind the starship, so that the ship and the laser beam were traveling in *opposite* directions as the cannon was fired? In that case, we must subtract the ship's velocity from that of the laser beam. Does that mean that the officer at the station would see the laser beam crawl through space at 0.01% of the usual velocity of light? What if the ship were traveling at exactly the speed of light while moving away from the direction in which it fired the beam? Would the light then have zero velocity in the frame of the space station? Can we even define a light beam with zero velocity?

Prior to 1887, nearly every scientist in the world would have proceeded in this manner. In Galilean relativity there is no absolute frame of rest, and all inertial frames are equivalent. The Galilean transformations were accepted at the time as the way to link observations made in one inertial frame with those made in another such frame. It was also known that Newton's laws of motion are invariant under Galilean transformations. However, the laws of electromagnetism, the Maxwell equations, are *not* invariant under the Galilean transformation. The Maxwell equations describe the behavior of fluctuating electric and magnetic fields, and those fluctuations depend specifically upon the speed of light. Hence a speed, the speed of light c , enters into Maxwell's equations in a fundamental way, and speed is not a quantity that is Galilean invariant. Maxwell himself believed that there should exist some special frame of reference in which his equations were correct as written; this would correspond to the frame in which the ether is at rest.

After 1887, the universe no longer seemed so simple. This was the year of the Michelson–Morley experiment, one of the crucial experiments that once in a great while turn our science upside down. The orbital speed of the Earth is large enough that Michelson and Morley's careful measurements should have easily determined the Earth's speed relative to the ether frame. Yet Michelson and Morley were unable to detect any evidence for motion with respect to this purported ether. In the absence of an ether to establish a frame for the speed of light, physicists were left with two unpalatable alternatives. The first possibility was that Maxwell's equations were incorrect, or perhaps that the physics of light was simply not the same in all inertial frames. The other alternative was that the Galilean transformation is invalid; but this would imply that something was amiss with Newton's mechanics. Yet Newtonian mechanics works so well for computing orbits; how could it possibly be wrong? On the other hand, the Maxwell equations were just as successful at explaining electromagnetism as the Newtonian equations were at explaining mechanics. How can we reconcile the invariance of one set of physical laws with the noninvariance of another?

*Maxwell's equations do not obey
Galilean relativity*

One of the first attempts to account for the null result of the Michelson–Morley experiment was made in 1889 by George F. Fitzgerald, who suggested that objects moving through the ether at velocity v were physically contracted in length according to

$$L(v) = L_0 \sqrt{1 - v^2/c^2}. \quad (7.1)$$

That is, a moving object would literally shrink by this amount in the direction of the motion through the ether. Such a contraction of the arm of the Michelson–Morley interferometer, in the direction parallel to the motion of the Earth, would shorten the travel distance for the light moving in that direction by precisely the amount needed to compensate for the change in the light propagation speed. Thus the round trip time would be equal for both arms of the apparatus, and no interference fringes would be seen. There was no fundamental theory to explain why objects would so contract; this was simply an *ad hoc* suggestion that reconciled the null result of the Michelson–Morley experiment with the existence of an ether. A hypothesis was put forward based on the recognition that intermolecular forces are electromagnetic in nature, so perhaps the very structure of matter was affected by motion through the ether. Yet this hypothesis seems very strange. How would an object be compressed? What if a living creature were to travel at a speed, relative to the ether, that was very close to that of light; would it be squeezed to death as v approached c ?

Many scientists rejected the Fitzgerald contraction, clinging instead to a more conservative interpretation. They struggled to explain the null result of the Michelson–Morley experiment as a consequence of a phenomenon called *ether drag*. If moving bodies dragged the ether along with them, then near the surface of the Earth no relative motion of Earth and ether could be detected. There was even, apparently, some experimental evidence for this; it had been known since the 1830s that the speed of light propagating through a moving fluid was different from its speed in a fluid at rest. When light travels through a medium, its speed is always less than its speed *in vacuo*, and depends upon the properties of the medium. For light traveling through a fluid such as water, some of the velocity of the fluid seemed to be imparted to the light, a phenomenon attributed to a partial entrainment of the ether by the fluid. This explanation preserved the old mechanical view of light, but at the expense of attributing to the ether even more strange properties, such as some kind of viscosity. If the Earth were dragging the ether, should it not lose energy, slow down in its orbit, and eventually fall into the Sun? This certainly had not occurred, nor was there any evidence for a systematic shrinking of the Earth’s orbit. Moreover, the ether-drag hypothesis predicted an effect on starlight as it entered the ether surrounding, and dragged by, the Earth; but no such effect was observed.

A bolder proposal was put forward by Ernst Mach. No motion relative to the ether was observed because there *was no* ether. An elegant experiment had been carried out that tested whether the ether existed.

The Fitzgerald contraction hypothesis

Ether drag

The ether rejected

The Lorentz transformation

The ether was not found; hence the ether theory was disproved. Accepting Mach's viewpoint still required the development of a new theory to replace the discredited ether theory. Several scientists took up the idea of the FitzGerald contraction, most prominently Hendrik Lorentz, and later Henri Poincaré and Joseph Larmor. They demonstrated that the Maxwell equations were invariant under a new kind of transformation law that makes use of this contraction. The new law, now known as the **Lorentz transformation**, contains a dilation, or slowing, of time, in addition to the **length contraction** proposed by FitzGerald. Thus it appeared that the Maxwell equations were invariant under the Lorentz transformation, while Newton's equations were invariant under the Galilean transformation. But neither of these transformations is arbitrary; they derive from basic ideas about the nature of space and time, so both cannot be correct. The Lorentz transformation, with its dilation of time and contraction of space, stands in direct opposition to something that was still regarded as more fundamental than Maxwell's equations: Newton's absolute time and space. Consequently, even the most eminent scientists of the day were reluctant to accept such a radical idea.

This unsatisfactory state persisted for several years. Equation (7.1), which is now most commonly called the Lorentz contraction¹, was adequate to explain the situation mathematically, but provided no physical insights. Poincaré seems to have been on the verge of realizing that the contraction was not absolute, but only, as we shall see, *relative*, yet he never developed a full theory. The new theory was brought forth in 1905 by a patent examiner in Bern, Switzerland, named Albert Einstein. The fact that Maxwell's equations did not seem to predict the same physics for observers in different inertial frames was disturbing to many scientists. What distinguished Einstein was that he found this *more* disturbing than the loss of Newtonian absolute space and time. Einstein had the audacity and courage to abandon Galilean relativity completely, and with it Newtonian mechanics, in order to preserve a property that he felt was most important. This property is embodied in his first relativity postulate:

- (1) *The laws of nature are the same in all inertial frames of reference.*

The postulates of relativity

This is called the **relativity principle**. In other words, there is no special frame of reference that is at rest with respect to some absolute space. All inertial frames of reference are completely equivalent. The relativity principle is also embodied in the earlier principles of Galilean relativity for Newtonian mechanics. Since Maxwell's equations do not preserve the relativity principle within Galilean relativity, Einstein chose to discard the latter rather than the former. To preserve the relativity principle for Maxwell's equations, we must adopt the second relativity postulate, which explains why the Michelson–Morley experiment produces a null result:

¹Sometimes it is called the Lorentz–FitzGerald contraction, in recognition of the scientist who first proposed it.

- (2) *The speed of light in the vacuum is the same in all inertial frames of reference.*

If the speed of light is the same for all inertial observers, then Maxwell's equations will be identical for all inertial observers, and no motion through the ether could ever be detected. Furthermore, if the ether is unobservable and has no detectable consequences, then we might as well consign the entire concept to the scientific junkheap. Light waves are not transmitted by a medium.²

Time dilation and length contraction

Despite their simplicity, the two relativity postulates contain remarkable, even incredible, consequences. Let us begin to explore some of the implications of these postulates, in particular the second. Our goal is to discover how the second relativity postulate implies the Lorentz transformation, and with it a change in the way we must view space and time. To do this, we will perform a “Gedankenexperiment,” or thought experiment.³ Relativity took hold during the days of train travel and has long been illustrated within that context; we shall hold to this tradition even though starships might really be more appropriate. As a first example, let us suppose that a train is traveling at a constant speed v , where we may consider v to be very close to the speed of light c . Inside the train, at the very center of one of the cars, a passenger switches on a light bulb at a certain time. We call such a discrete occurrence in space and time an **event**. To the passenger on the train, the light rays move at the speed of light and hit the front and rear of the car at the same time. There is nothing unusual about this.

Thought experiments reveal the consequences of the relativity postulates

Now consider the point of view of a train robber who is sitting on his motionless horse, just beyond the track, when the event of illuminating the bulb occurs. By the second postulate of relativity, light must move at the same speed c for the robber, even if that light is emitted from a moving bulb. The speed of the train is *not* added to or subtracted from the speed of the light the robber observes. Since the train is moving, however, the robber will observe the light to strike the rear of the car before it hits the front. The passenger judges the two events, the light's striking of the rear of the car and the striking of the front of the car, to be simultaneous; but for the robber, these two events occur at different times. Thus the property of **simultaneity** is not preserved for the two different observers. The difference in the passenger's and the robber's

Simultaneity is a relative concept

²Nevertheless, the word “ether” survives to this day in colloquial use. References to radio and TV signals moving through the ether are still common, and the Ethernet links computers.

³The term “thought experiment” can lead to confusion among those unfamiliar with the concept. A thought experiment refers to the deductive process of predicting the outcome of a specific experiment using the general principles of a theory such as special relativity. While these experiments may be difficult to carry out in practice, there is nothing in principle that prevents it. Many such relativity experiments have actually been performed.

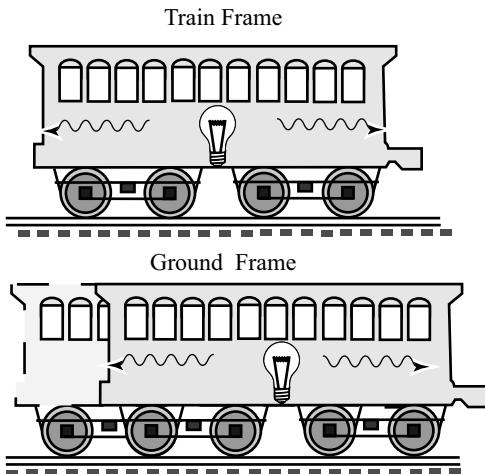


Fig. 7.2 A light flash goes off in the center of a moving train. In the train's frame, the light hits the front and back of the car simultaneously. In the ground frame, the train is moving with velocity v , so the light strikes the rear of the car before reaching the front. Two events that are simultaneous in one frame are *not* simultaneous in another frame.

descriptions of what happens is not due to light delay effects. Any delay due to the time required for light to travel some specified distance to an observer can be, and is, accounted for in describing what happens. Instead, time itself is somehow different for the two observers.

Now suppose that the passenger and the robber agree to perform another experiment, in which the passenger sets off a flashbulb on the floor of the train. The light travels to the ceiling and hits a mirror, which reflects it back to the bulb. Both observers measure the time for the light to make the round trip from the bulb to the mirror and back. For the passenger, this is straightforward; the height of the car is H , so the round trip travel time is simply $\Delta t_P = 2H/c$. It is simple for the robber as well, but for him the entire train moves some distance in the time it takes for the light to reach the ceiling and to return. Let the length of this angled path, from the floor to the ceiling, be denoted by d . By the second relativity postulate, the speed of light is the same in the robber's frame as in the passenger's,⁴ so the travel time he will measure is $\Delta t_R = 2d/c$. Since d is greater than H , Δt_R must be *greater than* Δt_P . We can regard this apparatus, in which light bounces from the ceiling and returns to its source, as a clock. One round trip is one tick of the clock. Our experiment shows that each tick of the clock takes longer, that is, time runs slower, for a clock located in an inertial frame that is moving with respect to some specified observer. This phenomenon is called **time dilation**.

How much slower is each tick of the clock, as measured in the frame relative to which it is moving? The robber remembers his geometry and

⁴Note that the constancy of the speed of light is the crucial assumption in special relativity. If Galilean relativity were applicable, the velocity of the light along the trajectory d would be the vector sum of its vertical and horizontal velocity components, and its speed would be the magnitude of whatever vector was thus obtained. But the second relativity postulate requires that the speed of light be the same for all observers.

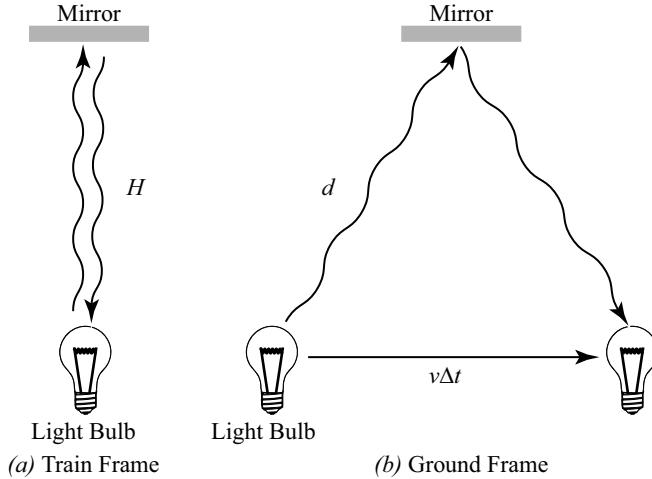


Fig. 7.3 Relativistic time dilation. The path of a flash of light traveling from a bulb to a mirror and back, as seen in (a) the rest frame of the train and (b) the rest frame of an observer on the ground who is watching the train pass by. In the ground frame the train moves to the right a distance $v\Delta t_R$ during the round trip of the light. The resulting light path is longer for the observer in the ground frame than for the observer in the train frame, but since the speed of light is the same in all frames, the time intervals measured in the two frames must differ.

uses the Pythagorean theorem to compute the distance d traveled by the light in his frame of reference:

$$d^2 = H^2 + \left(\frac{1}{2} v \Delta t_R \right)^2. \quad (7.2)$$

Recall that in the robber's frame, $\Delta t_R = 2d/c$, and in the passenger's frame $\Delta t_P = 2H/c$, so we can eliminate d and H to obtain a quadratic equation

$$\frac{1}{4}(c\Delta t_R)^2 = \frac{1}{4}(c\Delta t_P)^2 + \frac{1}{4}v^2(\Delta t_R)^2. \quad (7.3)$$

Working through the algebra leads us to the result

Time dilation

$$\Delta t_R = \frac{\Delta t_P}{(1 - v^2/c^2)^{1/2}}. \quad (7.4)$$

Since we have assumed that v is less than c , it follows that $1-v^2/c^2$ is less than 1, and therefore Δt_R is greater than Δt_P . This is the mathematical expression of the assertion that the light travel time measured by the robber is larger than the light travel time measured by the passenger. Thus, one tick of our bulb-and-mirror clock will be longer for a moving clock than the same tick will be for a clock at rest.

Does this result follow only because we have constructed an unusual clock with mirrors and light beams? What if we used an ordinary clock that did not involve light? But what is an “ordinary” clock? Suppose we used an atomic clock to measure the time interval between the departure and the return of the flash; would that make a difference? A clock is just a physical process with a regular periodic behavior. The details of the clock’s construction are irrelevant; no matter how we choose to measure the time interval, we shall always find that an observer moving with respect to the clock will observe that the interval for one tick is longer than it is for one tick in the clock’s own rest frame. Thus a given inertial observer will observe that any clock at rest with respect to his

Time dilation is a property of time, not of clocks

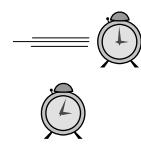


Fig. 7.4 A moving clock runs slow relative to one at rest.

The boost factor

frame will run at the *fastest* rate. All clocks moving with respect to some particular observer run slow relative to that observer's rest frame. Put more succinctly, "moving clocks run slow." We call the time measured by a clock that is at rest with respect to a specific inertial observer the **proper time**.

In equation (7.4) we found that the relationship between the time intervals in the two frames contained the factor

$$\frac{1}{(1 - v^2/c^2)^{1/2}} \equiv \Gamma. \quad (7.5)$$

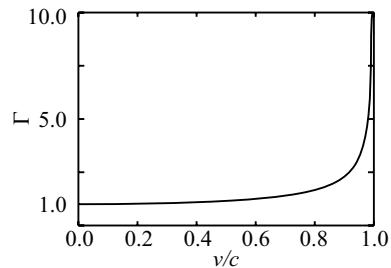


Fig. 7.5 The boost factor Γ as a function of speed, where speed is expressed as a fraction of the speed of light. The boost factor is 1 when $v = 0$ and becomes infinite when v is equal to c .

The reciprocal of this factor appeared in equation (7.1), the length contraction. We are now beginning to understand its significance. It does not tell us anything about a physical contraction of moving matter, but rather describes the way in which space and time are related for observers who are moving with respect to one another. It is often called the **boost factor** between two inertial frames. As an example, let us suppose that the train has a boost factor of $\Gamma = 2$, relative to the frame of the robber; this corresponds to a velocity of about $0.87c$. Suppose that the passenger's clock measures an interval of 30 seconds between some sequence of two events occurring on the train, such as the entrance of another passenger into the car and his exit on his way to the dining car. According to the robber, the interval between these two events is 60 seconds.

Now suppose it is the robber who has the bulb-and-mirror apparatus, and the passenger measures the interval between the flash and its return to the source. By assumption, the train is an inertial frame, so with respect to the train it is the passenger who is at rest, and the robber who is moving with velocity $-v$. The passenger must observe that his own frame is perfectly normal, and his clocks run at their usual rate, but the *robber's* clock is slow. This consequence is required by postulate (1) above; both observers must obtain the equivalent result when performing such an experiment. If the passenger were to observe that the robber's clock ran faster, then when the two of them compared their results they would be able to agree that it was the train that was actually moving. But there is no absolute frame of rest in special relativity, and therefore no inertial frame is *really* moving. Any two inertial frames are equivalent, and both must measure the same boost factor between them. This is the **principle of reciprocity**.

If time intervals are not invariant, what about space intervals? We measure time by the ticking of clocks, that is, by counting the number of cycles of some repetitive phenomenon. In our example above, we used a pulse of light bouncing back and forth between mirrors set a fixed distance apart. How do we measure separations in space? We do so by comparing the length of something to that of a fixed standard, a ruler, at some specific time. More explicitly, if we wish to measure the length of an object, we hold our ruler against it such that one end of the object lines up with one fiducial mark on our ruler, and *at the same time* the other end of the object lines up with another fiducial mark. Then we

The principle of reciprocity: frame A sees frame B as frame B sees frame A

count the number of marks between the two ends. Another method of measuring a length, such as the length of a jogging trail, is to travel from one end of the object to the other at a constant, known speed, and measure how long the transit takes. To derive the transformation of space intervals, let us return to the train and train robber, and consider two telegraph poles beside the tracks. The robber wishes to measure the distance between the poles, and to do so he will make use of the train, which is traveling at known speed v . The robber simply measures the time required for the front edge of the train to pass from the first pole to the second, Δt_R ; he thus determines that the distance between the poles must be $\Delta x_R = v\Delta t_R$. Now suppose the passenger on the train decides to measure the distance between the same two telegraph poles, which are moving with respect to him at speed v . Both the passenger and the robber agree on the relative speed v , as they must if inertial frames are to be equivalent. The passenger uses a similar timing technique of noting when the first, and then the second, pole passes the edge of his window; he measures a time interval Δt_P as the time between the passage of the first pole and the second. Thus the distance between the poles is $\Delta x_P = v\Delta t_P$. We have already solved for Δt_R in terms of Δt_P ; hence we can obtain

$$\frac{\Delta x_P}{\Delta x_R} = \frac{\Delta t_P}{\Delta t_R} = (1 - v^2/c^2)^{1/2} \quad (7.6)$$

or

$$\Delta x_P = \Delta x_R (1 - v^2/c^2)^{1/2} \quad (7.7)$$

which is exactly the Lorentz–FitzGerald contraction, equation (7.1). The passenger measures the distance between the moving poles to be shorter than the distance measured by the robber, in whose frame the poles are at rest. The distance measured by the passenger is specified by the self-same factor that was first proposed as an *ad hoc* explanation for some unexpected experimental results. Now it appears naturally and elegantly from the two fundamental postulates of special relativity.

Reciprocity applies to length measurements just as to time measurements. If the robber measures the length of one of the cars of the train, he will find this length to be shorter than that measured by the passenger, who is located within the rest frame of the car. Specifically, if the boost factor of the train is again $\Gamma = 2$, and the length of the car is 10 meters as measured by the passenger, then the robber will observe its length to be 5 meters. The result always depends on who is doing the measurement, as well as the relative speed of the object that is being measured. Similar to the proper time, we can define the **proper length** to be the length of an object as measured in its own rest frame. The proper length of an object is always the *largest* possible. In its own rest frame, the length of a meter stick always has the expected length of one meter. Meter sticks rushing past at any velocity are shorter than one meter.

So far we have discussed only the measurement of a distance that is parallel to the direction of motion. The length contraction occurs only

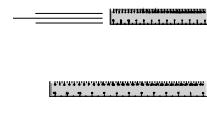


Fig. 7.6 The moving ruler is contracted relative to the one at rest.

Length contraction

Length contraction applies only along the direction of relative motion

along that direction, never in the direction perpendicular to the motion. To understand this, first recall that FitzGerald's proposal to explain the Michelson–Morley experiment invoked contraction only along the direction of the motion of the apparatus. It is easy to see that the situation would not have been clarified if both arms of the apparatus contracted by the same amount. As another illustration, let us apply the principle of reciprocity, and consider an argument in terms of the relativistic train. Suppose the train is 4 meters tall, and is traveling at such a speed that its boost factor is 2; it approaches a tunnel that has a 5-meter clearance. If there were a length contraction in the vertical direction, the robber, who is at rest with respect to the tunnel, would measure the height of the train as 2 meters, and would thus conclude that the train will easily fit into the tunnel. By reciprocity, however, the engineer must observe that the tunnel has only a 2.5-meter clearance, while the train is still 4 meters tall in its own rest frame; the train will be wrecked. But there cannot be a wreck in one frame and not in the other. Hence there can be no contraction along directions that are not in relative motion. Only that component of the relative velocity that is along the direction of the relative motion enters into the length contraction.

Special relativity and the Lorentz transformation lead to a quite unexpected view of the nature of time and space intervals. Why did scientists not notice the need for special relativity until near the end of the 19th century? Why did Newtonian mechanics and Galilean relativity seem to work so well for two hundred years? Let us compute the boost factor for one of the fastest macroscopic motions that anyone might personally experience on the Earth, the flight of a supersonic airplane such as the Concorde. The Concorde has a length of approximately 60 meters and when it was in service it flew, at top speed, at about twice the speed of sound. For this velocity, the boost factor is $\Gamma = 1.000000000002$, which makes its Lorentz-contracted length approximately 10^{-8} cm less than its rest length. This difference is about the diameter of an atom. For Earthly motions and speeds, the stuff of our everyday experiences, the effects due to the Lorentz transformation cannot be detected.⁵ When v/c is small, Γ is very close to one, and the Lorentz transformation reduces to the familiar Galilean transformation.

Relativity's effects are negligible at ordinary speeds

The meaning of the Lorentz transformation

Let us pause for a moment and reconsider these conclusions. Equations (7.4) and (7.7) together give the Lorentz transformation, the formulae for relating measurements in one inertial frame to those in another in-

⁵Atomic clocks, the only timepieces capable of measuring extremely tiny time differences, have been placed in airplanes and flown around the world. The cumulative effect due to time dilation, relative to an identical stay-at-home atomic clock, has been measured, and found to agree with the predictions of special and general relativity.

ertial frame. Length contraction and time dilation demonstrate that length and time intervals are different for observers moving with different velocities. This may seem disturbing; we are accustomed to thinking of space and time in Newtonian terms, as absolute, unvarying, and universal. Now they seem to depend on how fast an observer is moving, relative to some other frame. But what *are* space and time? How do we measure them? We have discussed briefly how distances can be measured. Since distance is defined to be the spatial separation between two points at some simultaneous point in time, it is clear that at a very basic level, time is tied into measurements of space. We measure time in terms of the number of occurrences of a regular physical process, such as the swinging of a pendulum. Today we may replace the pendulum with the vibration of a quartz crystal or the oscillation of a cesium atom, but the concept is the same. These are all physical processes, and we find that we define our concepts of space and time intervals in comparison with standard physical processes.

Since our measurement of time is so closely related to the behavior of physical systems, all means of measuring time within a given frame must give consistent results. Readers of science fiction occasionally encounter characters who are aware that time is running at some strange rate because they see their own clocks running slowly, or even, if it is a time-travel story, running backwards. But this is clearly absurd. If time slowed down (or stopped or reversed, whatever that means), then all physical processes would behave the same way, including the pulse of a heart, the speed of thoughts, the swing of the pendulum, or anything else by which the passage of time might be determined. The rate at which physical processes occur gives us our measure of time, and if all those rates changed together, an observer could not notice it. Try to imagine a means of measuring time that does not involve some periodic physical process!

Modern physics has shown that physical processes depend on the interaction of fundamental forces at a very basic level. The most important forces for everyday events are gravity and electromagnetism. In our modern view of physics, these long-range forces result from the exchange of massless particles: the graviton for gravity, and the photon for electromagnetism. Massless particles move at the speed of light. Gravitation and electromagnetism, therefore, propagate their forces through space at the speed we call the speed of light. Light, *per se*, is not particularly important here. We could just as well call it the speed of gravity, but few would understand what we meant. It is the speed attained by any massless particle, and hence the speed at which long-range forces propagate. This argument demonstrates that it was correct for Einstein to put the constancy of light ahead of the invariance of individual time and space measurements. For physical processes, the exchange of the particles that produce forces has ultimate importance. Moreover, the crucial distinguishing factor of special relativity is not so much the speed of light, as it is the existence of a finite speed of propagation of forces. Any such finite speed limit would result in a transformation law like the Lorentz

The significance of the speed of light

transformation; conversely, we may regard the Galilean transformation as that which would hold in a universe in which forces propagated with infinite speed. Thus special relativity does not stand in isolation, but is linked in a very profound way with the laws of nature.

The finite speed of propagation of the gravitational force solves one further nagging problem with Newtonian mechanics. Newton himself was somewhat disturbed by the instantaneous action at a distance that was implied by his gravitation law. Maxwell's equations described the transmission of electromagnetic force, but at a finite propagation speed. What happens to electromagnetism if we let the speed of light go to infinity? If c became infinite, the term v^2/c^2 would always be zero for any finite velocity, and, as we have stated, the Lorentz transformation reduces to the Galilean transformation. This demonstrates explicitly that the Galilean transformation is appropriate for instantaneous force transmission, such as is implied by Newton's law of gravity, whereas the Lorentz transformation is appropriate for finite speeds of force propagation. Framed in this way, perhaps Newton himself would agree that his gravity law must be modified in order that it will behave in a manner similar to that of Maxwell's equations. However, Einstein's special relativity alone does not accomplish this task. It is the general theory of relativity that reformulates the theory of gravity to include, among other features, a finite speed of propagation.

More transforms

Velocity addition in special relativity

This chapter began by contemplating what might be seen from a spaceship that shot a laser beam into space. Although we are now in a position to describe what happens in that situation, let us return instead to the simpler case of the train traveling down the tracks at speed v . The train passenger throws a ball down the aisle toward the front of the car. The speed of the ball in the frame of the train is v_b . The Galilean transformation would tell us that the speed of the ball in the robber's frame is given by the sum of its speed in the passenger's frame, plus the speed of the train, $v + v_b$. (If the passenger threw the ball toward the back of the car, the Galilean transformation would give us a speed in the robber's frame of the train speed minus the ball speed.) But we have discovered that the Galilean transformation is not correct. We also know from the postulate of relativity that the speed of light must be the same in all frames; hence we require a formula that yields, schematically, $\alpha(v + c) = c$, where α is a factor yet to be determined. From the Lorentz transformation, it is possible to work out exactly how velocities must add. We shall omit the details and merely present the *relativistic* velocity-addition formula

$$v_r = \frac{v_b + v}{1 + v_b v / c^2}, \quad (7.8)$$

where v is the relative velocity of train and robber, v_b is the velocity of the ball in the train frame, and v_r is the velocity of the ball as seen by

the robber. This equation has the desired property that if we replace the ball with a light beam and set $v_b = c$, we find that the robber also measures $v_r = c$. Hence, as required, the speed of light is the same in both inertial frames. Notice that this law also states that we cannot hope to achieve superluminal (faster than light) speeds with a hypercannon mounted on a relativistic spaceship. No matter how close two (sublight) speeds are to the speed of light, they can never add to a speed greater than that of light *in vacuo*.

Even though the speed of light is the ultimate speed limit, there is still a great difference between, for example, $0.9c$ and $0.99c$. Suppose a starship has on board a particle accelerator that can eject a beam of matter at $0.9c$, as measured in its own rest frame. If the spaceship wishes to project a beam at a speed of $0.99c$, as measured by an observer on the space station, at what speed must the spaceship travel, relative to the space station? The velocity-addition formula, equation (7.8), tells us that the spaceship must have a relative velocity of a little under $0.9c$ in order for the beam to reach $0.99c$. In other words, $0.99c$ is about twice as fast as $0.9c$! If we compute the boost factors, we find that $\Gamma = 2.3$ for $0.9c$, while $\Gamma = 7.1$ for $0.99c$. In terms of velocity increase, it is about as difficult to increase from $0.9c$ to $0.99c$ as it is to go from zero velocity to $0.9c$.

The relativistic addition of velocity also explains the apparent experimental observation of ether dragging by a moving fluid. Nineteenth-century scientists had observed that when light was propagating within a moving fluid it seemed to travel faster than when it propagated within a fluid at rest; this was often cited as experimental support for the dragging of ether before Einstein developed special relativity. In view of our present knowledge, we can see that what was actually observed was the relativistic addition law, although this was not realized until fully two years after the publication of Einstein's paper on special relativity. Light propagates through a fluid at a speed less than c ; the exact speed is a function of the index of refraction of the substance. If the fluid, in turn, is moving with respect to the experimenter, the observed speed of the light in the frame of the experimenter is given by equation (7.8). It could be argued that special relativity was confirmed experimentally before it was even conceived, but of course the correct interpretation could not be seen until Einstein was able to break through the prevailing patterns of thought.

The *speed* of light is the same in all inertial frames, but this does not imply that all inertial observers see light in exactly the same way. The frequency of a light wave specifies how many times per second that the electromagnetic field oscillates, and hence frequency is a kind of clock; thus it is affected by time dilation. The clock of a moving source runs slow; thus the frequency of a moving source must be reduced, as measured by the receiver. This effect is independent of the classical Doppler effect; not only is there a bunching up (approach) or stretching out (recession) of the wave crests owing to the relative motion of source and receiver, but there is also a relativistic correction due to time dilatation.

Experimental confirmation of the relativistic velocity addition law

The relativistic Doppler effect

tion. Recalling the definition of redshift, equation (4.1), the relativistic Doppler formula is given by

$$z + 1 = \sqrt{\frac{1 + v/c}{1 - v/c}}. \quad (7.9)$$

A relativistic Doppler effect also occurs in the direction perpendicular to the relative motion, exclusively as a consequence of time dilation; because of this effect the frequency of the light is reduced even if it is approaching along a direction perpendicular to the motion of the emitter. The transverse Doppler effect is very small; it is practically unobservable for most motions. However, a few astronomical objects do exhibit relativistic motions with detectable transverse Doppler shifts. One such object is the remarkable SS433, a star in the constellation Aquila, about 16,000 lightyears from Earth. In radio images, the star evinces jets of gas extending from the main source. SS433 is a binary system consisting of a normal star and a compact object, most likely a neutron star; its location within an ancient supernova remnant supports this scenario. Gas from the normal star is drawn into an accretion disk around the neutron star, and some of it is squirted at relativistic velocities in two oppositely directed jets, along the axis of the neutron star. The jets wobble, and when the beam is directly perpendicular to our line of sight a purely relativistic, transverse Doppler shift can be observed in the spectrum, corresponding to gas speeds of about one quarter of the speed of light. No matter how extreme the behavior, somewhere in the observable universe there is usually some object that demonstrates it.

The strange velocity addition rule of special relativity hints at another important consequence besides the intermingling of space and time, time dilation, and length contraction. It leads us to what is perhaps the most famous equation in history, $E = mc^2$. But what does this renowned equation *mean*, and how does it fit into relativity theory? To find the answer, first we must specify what we mean by energy. We have previously defined energy as “the capacity to do work.” In the Newtonian universe, energy is not created or destroyed, but is only transformed from one form to another. Similarly, there is a separate conservation law for matter; matter is neither created nor destroyed. One of the most important forms of energy is *kinetic energy*, or the energy of motion. In Newtonian mechanics, it can be shown that the kinetic energy of a particle is given by

$$E_k = \frac{1}{2}mv^2, \quad (7.10)$$

where m is the mass of the particle and v is its speed. The Newtonian kinetic energy of a particle at rest is zero. The Einsteinian equation is the relativistic generalization of this concept of kinetic energy. Einstein’s equation is more correctly written as

$$E = \Gamma m_0 c^2, \quad (7.11)$$

where Γ is our new acquaintance, the boost factor, and m_0 is the *rest mass* of the particle, that is, its mass as measured in its own rest frame.

Relativistic energy

Notice that, since the boost factor is 1 for a particle at rest, this definition of energy does not vanish for $v = 0$. Thus relativity shows that there is a **rest energy**, given by m_0c^2 , associated with every massive particle. As the speed of the particle increases, its energy also increases. For velocities small compared to c , it is possible to show that the relativistic energy equation reduces to

$$E = m_0c^2 + \frac{1}{2}mv^2 + \text{additional terms.} \quad (7.12)$$

For nearly all motions in the macroscopic world, the additional terms are very, very small and we recover the Newtonian law, with the addition of the new concept of the rest energy. At the other extreme, as the speed increases and begins to approach that of light, the relativistic energy (7.11) becomes very large, much larger than the simple Newtonian rule would predict; it is arbitrarily large for speeds arbitrarily close to the speed of light.

As an example, consider how much energy would be required to accelerate 1 kg of matter to $0.87c$, for which $\Gamma = 2$. In order to compute the relativistic *kinetic* energy, we must subtract from the total energy, as given by equation (7.11), the rest energy specified by m_0c^2 . Carrying out this procedure, taking care to keep our units consistent, we obtain a result of 9×10^{16} J of kinetic energy. In units that might be more familiar, this is 3×10^{10} kilowatt-hours, or about 20 megatons TNT equivalent. In other words, it would require all the energy released by a very large thermonuclear bomb in order to accelerate just one kilogram of matter to near the speed of light. This is a serious limitation on our ability to boost anything, even elementary particles such as protons, to speeds close to that of light. At accelerator laboratories around the world, scientists do succeed in accelerating protons and other particles to relativistic speeds. It is no coincidence that thick power lines lead onto the grounds of these accelerators!

Mass as a form of energy

The rest energy is an interesting concept. Does it represent some irreducible amount of mass that is always conserved, or does it mean that energy and mass are truly equivalent and can be transformed into one another? When Einstein wrote down his famous equation, no experimental evidence existed to decide that issue; but he chose to interpret his equation boldly, asserting that the equals sign meant just that, equality. Mass can be converted into energy, and energy into mass. The rest energy can be interpreted as the energy due to the inertial mass; in this view, inertial mass is itself just another form of energy. Subsequent events have proven that this assertion is correct. If even a tiny fraction of the rest energy of a particle is converted to another form of energy by some means, the yield can be enormous. For most people, nuclear weapons most dramatically illustrate the principles of special relativity. Ironically, nuclear reactions are relatively ineffective at extracting rest energy; they just happen to be the best mechanisms available on Earth. Nuclear reactions, both fission and fusion, convert about 1% of the rest mass involved into other forms of energy. However, one percent of the

rest energy of even a few kilograms of fuel is an enormous amount of energy. Although weapons may be the most familiar application of nuclear reactions, fusion in particular is of utmost importance to humans; fusion reactions occurring at the core of the Sun are ultimately responsible for the existence of life on Earth. The Sun has so much mass that one percent conversion provides power for tens of billions of years at its current luminosity, so this small efficiency is adequate for our needs. The most efficient process possible is matter–antimatter annihilation, in which a particle and its antiparticle are both converted *completely* into energy.⁶

Sometimes it is believed that equation (7.11) applies only to such exotic reactions as matter–antimatter annihilation or nuclear reactions. In fact, *any* release of energy, including mundane ones such as chemical processes, results in a change of mass. In the nonrelativistic world this change in mass is unmeasurably small; yet it occurs. Conversely, energy can be converted into matter; this is a stunning illustration of the equivalence of mass and energy, and hence of the special theory of relativity. Just as we can no longer think of space and time individually, we must not think of energy and matter as distinct quantities. Both are revealed to be two aspects of the same entity, mass-energy, and it is mass-energy that is conserved.

High-energy subatomic particles exhibit relativistic properties

Examples of relativistically significant energies occurring near Earth are not easy to find. One example is given by particle accelerators, in which high-speed particles are slammed together. At their large boost factors, they have energy to spare to produce a shower of particles and antiparticles. However, at least one everyday relativistic motion occurs regularly: the flight of muons through the atmosphere. Muons, a kind of heavy lepton with an extremely short half-life, are created high in the upper atmosphere when cosmic rays, which are actually high-energy particles from the Sun and other sources, collide with atoms. In the collision, some of the kinetic energy of the impinging particle is converted into matter, such as muons. In its own rest frame, the muon decays with a mean life expectancy of only 2 microseconds. Atmospheric events produce muons traveling at typical speeds of $0.99995c$, corresponding to a boost factor of about 100. Even at this speed, without relativistic effects a typical muon would travel only 600 meters over its half-life, and practically none would ever reach the surface of the Earth. Yet the surface is constantly bombarded by relativistic muons. In the frame of the Earth, the muon exists 200 microseconds, in which time it travels 60 kilometers! The arriving muons also have a relativistic mass 100 times greater than their rest mass.

⁶The writers of the television show *Star Trek* chose matter–antimatter engines to power their starship for good reason. With 100% conversion of mass to energy, this would be the most efficient engine possible. One practical problem with such engines is obtaining antimatter for fuel. As far as we know, the universe is composed almost entirely of ordinary matter.

Space-time

We have discovered that time intervals, space separations, and simultaneity are not absolute, but depend upon the frame of reference of the observer. Suddenly, the orderly Newtonian universe has been replaced by a much more unruly one, where space and time are relative to the observer, and where a measurement of time depends on space and *vice versa*. Most of us grow up thinking of space and time as absolute, distinguishable quantities; now we find that they somehow intermingle. We must no longer think of space and time as separate entities; rather than three space dimensions and one time dimension, our new view of the universe is a four-dimensional **space-time**. The special-relativistic universe can be represented by a mathematical entity called **Minkowskian space-time**, named in honor of Hermann Minkowski, who first formulated its properties in 1908. Minkowskian space-time provides a way to unify the mixing of time and space, as given by the length contraction and time dilation, into one four-dimensional structure. Unfortunately, it is difficult to think in four dimensions, and impossible to visualize any four-dimensional object. However, a very useful device for representing occurrences in this space-time is the **space-time diagram**. Usually we simplify matters by plotting only one space dimension, since we cannot show more than two of them anyway. We cannot draw, or even model on a tabletop, four mutually perpendicular axes, but one space dimension is generally adequate for the purpose of understanding a physical process. The remaining axis is labeled as time. Together, the time and space coordinates represent some inertial reference frame. The position x and time t of an object can be plotted on such a graph, tracing out a continuous curve on a space-time diagram. Any point on such a curve is an event, and the curve itself is called a **worldline**. (Strictly speaking, the worldline describes only the path of a point particle; any larger object is a collection of points, each moving along a worldline, so the object as a whole traces out a *world tube*. The distinction is not very important here.) On such a diagram, a straight line corresponds to an object moving with a constant velocity, that is, an inertial observer; the slope of the line is proportional to the speed of the observer. If the worldline curves, that is, its slope changes, then the velocity is changing, and the curve corresponds to a noninertial observer.

It is usual to calibrate the time variable on space-time diagrams as ct , so that both axes are labeled with the same units. Any observer at rest with respect to these coordinates traces a worldline that is vertical, that is, it remains at a constant space position. A light beam follows a worldline given by $x = ct$; with our units convention, such a path is a straight line at an angle of 45° . A massive observer moving inertially with respect to the plotted coordinate system follows a worldline that is also straight, but always forms an angle greater than 45° with respect to the space axis; that is, the velocity is less than c . If the slope of the line connecting any two events on the diagram is less than 45° , then those events could be connected only by moving faster than light.

The concept of space-time

A standard convention: light follows a 45 degree path through space-time

Fig. 7.7 A space-time diagram showing three worldlines. A point on a worldline is called an event. In the inertial frame of (x, t) , any straight worldline corresponds to inertial motion. The perfectly vertical line corresponds to an observer at rest with respect to the coordinates. A worldline that curves corresponds to an accelerated observer.

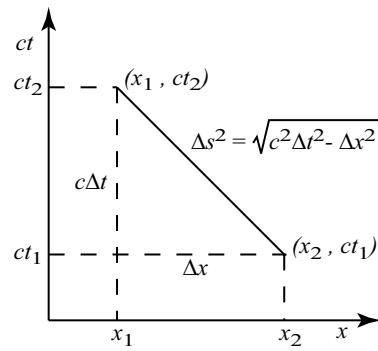
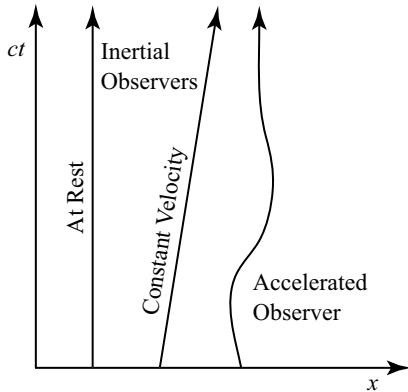


Fig. 7.8 The rule for the space-time interval between events located at (x_1, ct_1) and (x_2, ct_2) . It differs from the Pythagorean rule (cf. Figure 6.5) by the presence of the negative sign between the space interval and the time interval. The space-time interval is invariant under the Lorentz transformation.

The space-time interval measures an invariant distance in space-time

If space-time were like our usual (x, y) space coordinates, we would be on familiar ground. Given two points in (x, y) coordinates, we could use the Pythagorean theorem to obtain the distance between them. Recall that the square of the distance along the hypotenuse of a right triangle is equal to the sum of the squares of the other two sides:

$$(\Delta r)^2 = (\Delta x)^2 + (\Delta y)^2. \quad (7.13)$$

This Pythagorean formula defines the shortest distance between two points on a flat surface, in terms of perpendicular coordinates x and y . But space-time is *not* like ordinary space. Given two events that lie close to one another on a worldline, how can we define a distance between them in Minkowskian space-time? Each of the two events occurs at points labeled by appropriate (x, ct) coordinates; the distance, or **space-time interval**, between them is defined by

$$\Delta s = \sqrt{(c\Delta t)^2 - (\Delta x)^2}. \quad (7.14)$$

The principal difference between the space-time distance formula and the Pythagorean formula is the negative sign between the time interval and the space interval. Time may be the fourth dimension, but it still differs from the three space dimensions. The properties of this space-time geometry are different from those of the ordinary *Euclidean space* to which we are accustomed.

The space-time interval derives its fundamental significance from the fact that it is *invariant* under the Lorentz transformation. Given two events, all observers can measure both the space and time intervals separating the events. We have learned that different observers will obtain different results for the individual space and time intervals, with the measurements related for inertial observers by the Lorentz transformation. However, when any observer combines his own values for the space and time intervals between two events into a space-time interval, as specified by equation (7.14), *all* observers will obtain the same result. This means that no matter what coordinates are used, or what an observer's velocity is with respect to the (x, ct) coordinates, all observers will obtain the same result for the value of the space-time interval. (This fact

is easily demonstrated by applying equations [7.4] and [7.7] directly to equation [7.14].) It is comforting to find that after we have thoroughly mixed up space and time, we finally have something that is invariant.

Is there a physical interpretation for the space-time interval? Suppose an inertial observer measures this interval using a clock at rest with respect to herself. In her own rest frame, she is always at the same location, so Δx is zero. All that remains is the time interval; hence the space-time interval corresponds to the *proper time*, which we have already encountered. Since it is invariant, *any* inertial observer can measure it, and each will obtain the same value, even if different observers may disagree about the component space and time intervals. In simplified terms, everyone will agree about how much time elapses on a given person's watch as that person travels from one event to another, although different observers may disagree on how that proper time interval compares to the time elapsed on their own watches.

The space-time interval measures proper time

Suppose we are interested in the space-time interval between two events along some arbitrary *curved* worldline. In that case, we cannot use equation (7.14) directly. In order to compute this quantity, we must consider a large number of pairs of events close together along the curve, calculate each space-time interval, and add them up.⁷ The space-time interval between two events depends on the path between them, which at first might seem quite remarkable. With some reflection, however, this ought not to seem so odd; after all, the distance from one's home to the nearest grocery store depends on the path chosen. Since the space-time interval equals the proper time along that worldline, we see that the elapsed proper time kept by two clocks that travel on different paths through space-time between the same two events can be different.

What is the space-time interval between two nearby events on the worldline of a light ray? Since the light ray always travels at speed c , for it $\Delta x = c\Delta t$, so the space-time interval is zero. But the sum of any number of zeroes is still zero; therefore, the space-time interval between *any* two events on the lightbeam's worldline is always zero, for any observer! This result shows us immediately that the space-time interval does not behave like the Pythagorean distance of Euclidean space. The Euclidean distance is always positive for any two distinct points, and is zero if and only if the points coincide. In contrast, the *square* of the space-time interval can be positive, negative, or zero, and we can use this fact to divide all space-time intervals into three classes. If the space-time interval, and its square, are positive, the interval is said to be **timelike** (that is, the time interval $c\Delta t$ is greater than the space interval Δx). Timelike worldlines describe paths that can be traversed by massive, physical particles, all of which must travel at less than the speed of light. If the square is negative, the interval is **spacelike**.⁸ If two

Intervals in space-time may be space-like, timelike, or lightlike

⁷The reader who has studied calculus might remember that the distance along a curve is found by integrating the differential arc lengths from the beginning to the endpoint. The space-time interval corresponds to the arc length in space-time.

⁸Numbers whose squares are negative make up the set of *imaginary* numbers. The imaginary numbers are probably less familiar than the real numbers, but they

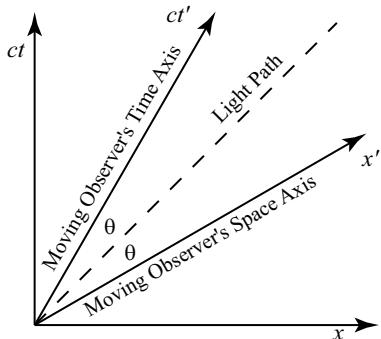


Fig. 7.9 Relationship between the space-time coordinates (x', ct') of an inertial observer to the coordinates of another observer (x, ct) with respect to whom the first observer is moving. The angles θ formed by the t' and x' axes with respect to the lightcone $x = ct$ are equal.

Time and space axes for a moving observer

events are separated by a spacelike interval, they cannot be connected either by a light ray or by the worldline of a particle traveling at a speed less than the speed of light.⁹ If the space-time interval is zero, it is said to be *null* or **lightlike**. Any particle that travels on a such a lightlike worldline must have zero mass, that is, the particle is *massless*. The converse is also true; a massless particle must always travel at the speed of light. Photons can be created or destroyed, but they cannot slow to less than c , the speed of light *in vacuo*. The speed of light itself may differ in different media; for instance, the speed of light in water is less than its speed in a vacuum. This occurs because the atoms or molecules of the substance absorb photons and then, after some interval of time, re-emit them. The photons themselves always travel at speed c between interactions.

To this point we have discussed observers who are at rest with respect to the inertial frame (x, ct) . What about an observer who is moving with respect to this inertial reference frame? How should the space and time axes of a moving observer be drawn onto our space-time diagram? First, construct an inertial set of coordinates (x, ct) and draw the space-time diagram. Suppose we label the moving observer's own coordinates by (x', ct') . The time axis is easy: it must lie along the observer's worldline, because an observer is always at rest in his own coordinate frame. The position of the space line is less obvious. Our first impulse might be to draw it perpendicular to the ct' axis. This is appropriate for an ordinary spatial coordinate system such as (x, y) , but in Minkowskian space-time the one thing that must be preserved is the constancy of the speed of light in all frames. In our (x, ct) frame, the lightlike worldline of a light beam runs halfway between the x and ct axes such that $\Delta t = \Delta x/c$. The same must be true for the (x', ct') frame. Hence we must draw the x' axis so that the light beam makes an equal angle with both the ct' and the x' axes, as shown in Figure 7.9. Perpendicular axes that do not form a 90° angle with respect to one another may appear strange, but here again space-time is not like our familiar flat-space Euclidean geometry. The canting of the (x', ct') axes occurs because it is impossible to represent Minkowskian space-time exactly on a flat sheet of paper. As an analogy, consider a drawing of a three-dimensional cube on two-dimensional paper. Even though the edges of a cube always form an angle of 90° where they meet in the three-dimensional space in which the cube exists, the drawing has line edges at different angles. A drawing of a three-dimensional cube on a two-dimensional sheet can only approximate the actual shape of the cube. Similarly, Minkowskian space-time can be represented only approximately on a two-dimensional, Euclidean paper.

represent a perfectly valid mathematical concept, useful in many tangible fields such as engineering. However, we shall have no need to make explicit use of imaginary numbers in this text.

⁹There has been some speculation about whether or not particles could exist that must always travel faster than the speed of light. Such particles are called *tachyons*, and if they existed they would have very strange properties indeed.

Some interesting features can be observed on Figure 7.9. Two events occur at the same place for a given observer if their spatial location is the same, in that observer's frame. For the observer at rest with respect to the (x, ct) frame, two events that occur at the same place will lie along a vertical line; that is, Δx between those events will be zero, though they will be separated by a time interval. The moving observer (the primed frame) will not see such events occur at the same point in space, however, but will see a $\Delta x'$ as well as a $\Delta t'$, in general. This conclusion is not remarkable; it holds in Galilean relativity as well. In four-dimensional space-time, we must extend this concept to include time. Two events are simultaneous if they occur at the same time but at different places; that is, simultaneous events have zero Δt in some frame. The stationary observer will note that simultaneous events lie along a line parallel to the x axis. But the moving observer will *not* observe those events to be simultaneous. In the frame of the moving observer, simultaneous events must lie along lines parallel to the x' axis, which is slanted with respect to the x axis. For events simultaneous in the primed frame, $\Delta t'$, but *not* Δt , is zero. Simultaneity is relative.

Is there anything about which all observers agree? Because the space-time interval is an invariant and any function, such as the square, of an invariant is also invariant, all observers will agree that a given interval is either timelike, spacelike, or lightlike. Let us draw a space-time diagram to illustrate this. First, draw the possible worldlines of a photon that passes through some particular event, labeled A . The photon can travel toward the left or right, but must always move with speed c , so its possible worldlines are two lines at angles of 45° to the horizontal. If we were to plot more spatial dimensions, the possible worldlines of the photon would lie on a cone, so this dividing surface is called the **lightcone**, or, sometimes, the *null cone*. The surface of the lightcone divides space-time into distinct regions for all observers. The region inside the cone corresponds to all events that are separated by timelike intervals from event A . The region outside the cone contains all those events separated from A by spacelike intervals. The half-cone below A is called the *past lightcone*. The past lightcone and the timelike region within it make up the **past** of event A . Similarly, the half-cone above A is called the *future lightcone* of A , and this half-cone and the timelike region that it encompasses comprise together the **future** of A . Events outside the lightcone of A are in the **elsewhere** of A . Given two events, B within the past lightcone of A , and C within its future lightcone, it can be shown that *all* observers will agree that B is in the past of A and C is in its future, although they may not agree about where and when within the past and future, respectively, these events occur. Thus all observers agree that B occurs before C . On the other hand, for an event D in the elsewhere of A , observers may disagree on the order of events A and D ; some may see that A occurs first, while others may observe that D happens first, and still others may regard the two events as simultaneous.

Simultaneous events are those that lie along the coordinate x -axis at constant coordinate time

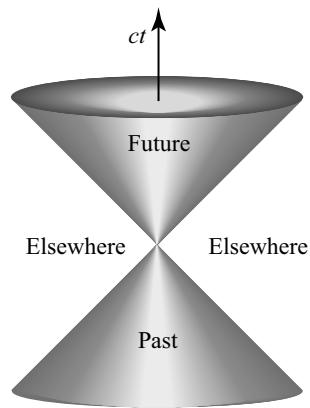


Fig. 7.10 When drawn with two spatial dimensions, the lightcone appears as a cone.

The lightcone divides space-time into three regions relative to a particular event

Relativity preserves the concept of cause and effect

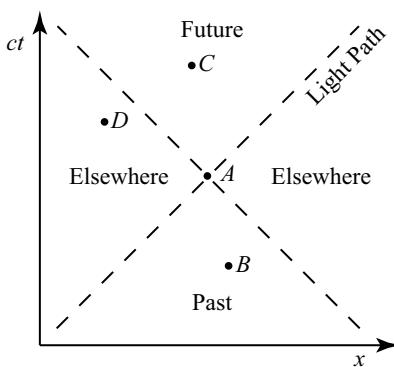


Fig. 7.11 The lightcone seen on a simple (x, ct) plot. A lightcone can be constructed at each event in a space-time. The lightcone divides space-time into those events in the past (event B), the future (event C), and elsewhere (event D) with respect to event A .

Looking outward into the universe is equivalent to looking into the past

If different observers do not agree on the ordering of events that are separated by spacelike intervals, does it follow that the principle of cause and effect is in jeopardy? Can we find a frame in which the lights go on before the switch is thrown? Most of us believe intuitively that a cause must always precede its effect. This has been formalized into the **principle of causality**, and it is one of the guiding principles of physics. It cannot be proven from any physical laws; it is, in some respects, one of the *axioms* of physics. But without it, we cannot make sense of the universe. Science is based on the belief that we *can* understand the universe; its success at this endeavor is ample demonstration of the power of its axioms. It is the principle of causality, ultimately, which asserts that no physical particle can travel faster than the speed of light. We already have ample evidence that there is something special about the speed of light. For one thing, the boost factor becomes infinite at that speed, and its reciprocal is zero; time dilation goes to infinity, while objects are Lorentz-contracted to zero length. Equation (7.11) shows that the relativistic mass of a particle is infinite at the speed of light, unless the rest mass of the particle is zero; this justifies our statement that photons are massless, since we know that photons have finite energy. (It also demonstrates, though it does not prove, the converse, that massless particles move at the speed of light.) These strange phenomena suggest that the speed of light sets the upper limit to speed in the observable universe. Yet it is causality that *requires* this to be so. The ordering of events is guaranteed only for timelike or lightlike space-time separations. If information could travel faster than the speed of light, that is, along spacelike worldlines, then the principle of causality could be violated; an effect could precede its cause in some frames. Events separated by a spacelike interval *cannot* be causally connected to one another. Therefore, no information, nor any physical particle that could carry it, can travel faster than the speed of light. Fortunately, the invariance of the space-time interval means that if B is the cause and C is its effect in one frame, then B always occurs before C in *any* frame of reference. Thus within special relativity the relationship between cause and effect is preserved; cause always precedes effect, for all observers.

The lightcone is a useful concept even in the nonrelativistic world. Looking up at twilight to see the first star of the evening means looking back along the past lightcone to the event at which the photons that form the image of the star left the surface of that star. Observing a distant star is equivalent to looking backwards in time. In fact, every image is a picture along the past lightcone, for photons bring us the information by which we see. For the things of this Earth, the time delay is generally of no significance. But when we seek to study the contents of the universe, those objects that are at farther and farther distances are seen as they were at earlier and earlier times in the past; we can never form a picture of the universe as it is now. In some respects, though, this is a benefit, for it means that as we look out through space we can see the history of the universe laid out before us.

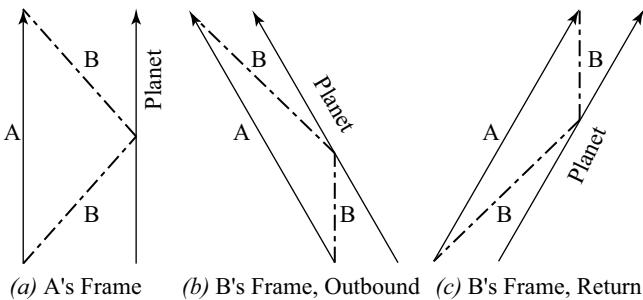


Fig. 7.12 Three space-time diagrams describing the twin paradox. Each diagram shows the point of view of a different inertial frame. (a) The twin paradox in the inertial frame of the stay-at-home twin *A*. (b) The twin paradox in the inertial frame of *B* during her outward voyage. (c) The twin paradox in the frame of *B* during her return trip. The essential asymmetry between *A* and *B* is that *B* does not remain in one inertial frame for the entire journey.

Some paradoxes of special relativity

The twin paradox

Our study of space-time diagrams will make it much simpler to understand one of the most famous paradoxes of special relativity, the so-called Twin Paradox. Andy and Betty are fraternal twins. Betty is chosen to go on the first mission to Alpha Centauri. She rides in a spaceship at nearly the speed of light, visits Alpha Centauri, then returns at nearly the speed of light. Andy stays on Earth and waits for her return. While Betty was traveling at relativistic speeds, her clocks, including her life processes, ran slow relative to Andy's frame; therefore, upon her return, she is younger than Andy. This seems like a straightforward solution, until we consider Betty's point of view. To Betty, the Earth receded at nearly the speed of light and Alpha Centauri approached. Subsequently, the Earth returned at nearly the speed of light. In her frame, her clock was always running normally. Since it was Andy who receded and approached at nearly the speed of light, *Andy* should be younger. Who is right?

The resolution lies with the realization that Betty started from rest, turned around at Alpha Centauri, and returned to a state of rest on Earth. Consequently, she must have accelerated at least four times on her trip. The accelerations mean that she did *not* remain in a single inertial reference frame, whereas Andy did remain in his inertial frame. It is only *inertial* frames that can be equivalent, so there is no paradox here; the accelerated twin changed inertial frames before returning to Andy's frame, which he never left. Betty, the traveling twin, is younger. If Andy had decided to join Betty after her departure, he would have had to hop onto the next mission in a spaceship fast enough to catch up to her. In this case, when he reached her ship they would find that Andy would now be younger than Betty. If Andy leaves a given inertial reference frame, hence experiencing an acceleration, and then rejoins the same inertial frame, which necessarily requires another acceleration, Andy will find that his clocks show less elapsed time than do the clocks of those who never left the original inertial frame.

Why is the traveling twin younger?

The stay-at-home twin remains in a single inertial frame

Space-time diagrams for the twin paradox

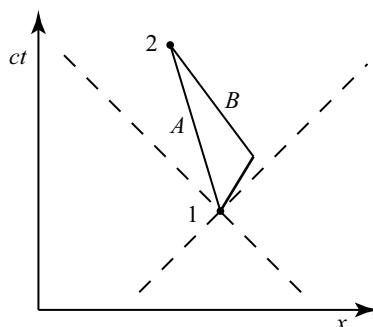


Fig. 7.13 Worldlines A and B connect events 1 and 2. The straight worldline, A, is inertial and experiences the maximum proper time. Accelerated worldlines, like B, experience less proper time. Clocks on such trajectories run slow relative to those on straight worldlines.

The traveling twin minimizes her elapsed proper time by traveling “close to the lightcone”

We can illustrate the Twin Paradox on a space-time diagram. First we must recognize that there are *three* inertial reference frames relevant to this problem: the first is the frame of the stay-at-home twin, Andy. (We will ignore the motions of the Earth, since they are small compared to the speed of light. We will also assume that Alpha Centauri is at rest with respect to the Earth, although this is not essential.) The second is the inertial reference frame of Betty while traveling to Alpha Centauri, and the third is her inertial reference frame during her return voyage. Figure 7.12 illustrates the round trip in all three of the inertial frames. Betty departs for Alpha Centauri, traveling at constant, very high, velocity, at Event 1. She turns around at Event 2, reversing her direction and returning at the same high, constant velocity. She arrives home at Event 3. Notice that in all three reference frames, Andy’s worldline is straight, indicating that he remains inertial for the entire trip. In contrast, Betty’s worldline is not straight in all three diagrams, but changes direction at event 2.

We learned from our discussion of space-time intervals that the elapsed proper time along any worldline is obtained by summing the space-time intervals along that worldline. Therefore, we should not be surprised that the proper times recorded by clocks traveling along different worldlines between the same two events (Betty’s departure and return) will be different. The general rule for such circumstances is that the *maximum* amount of proper time between any two events is that recorded by a clock that follows the straight line through space-time between those two events; that is, the clock which remains in one inertial, constant-velocity frame (see Figure 7.13). This means that in Minkowskian space-time, the *longest* time between any two events is a straight line! The fact that the straight line is a maximum, rather than a minimum, is another consequence of the negative sign in the space-time interval, and another way in which relativity can confuse us. Euclidean space and Minkowskian space-time have different properties.

If the maximum proper time is obtained by the inertial clock moving along the straight line in space-time, which clock would show the minimum time? If you wish to record zero time between two events, there is only one way to do it: follow the light beam. A beam of light sent out into space and bounced back to Earth would follow a noninertial, yet still lightlike worldline, and the space-time interval along any lightlike path, accelerated or not, is always zero. Objects with nonzero mass, such as Betty, cannot travel at the speed of light, but Betty can minimize her proper time between two events by traveling as close as she can to the lightcone. In all three frames of the Twin Paradox, only the worldline of Andy is a straight line through space-time. Betty travels from event 1 to event 3 by a noninertial route close to the lightcone. Hence her clock reads less elapsed proper time than does Andy’s, and the faster Betty travels, the smaller her elapsed proper time. There is no paradox once we understand this.

It is occasionally claimed that the Twin Paradox requires general relativity for its resolution. This is completely incorrect. For some reason, it

is a common misconception that special relativity cannot accommodate acceleration, that general relativity is required to deal with it. As we shall learn in the next chapter, general relativity is a theory of gravitation, and has nothing particular to say about acceleration *per se* that cannot be treated within special relativity. There is a formula for relativistic acceleration, as well as a relativistic generalization of Newton's second law. We have not shown them because they are rather advanced, not often used, and not essential to our story here. But they exist. It is more difficult to carry out the calculation of proper time when acceleration occurs, almost always requiring calculus and some rather messy manipulations. But if we know the accelerations, we can do it. In our discussion above we simply assumed that Betty's intervals of acceleration during her journey were small, and ignored the small corrections that would be obtained by integrating carefully along those curved paths in her worldline. Regardless, we find that if Andy stays home, Betty is unequivocally the younger twin upon her return.

The clock paradox

Let us test our new-found understanding of special relativity with another example, illustrated in Figure 7.14. A succession of outposts is established in space. Each has a master clock, and all the space stations, and their clocks, are at rest with respect to one another. By a clever arrangement, we switch on all the clocks at the same time; that is, the events consisting of the starting of the clocks are simultaneous in the frame of the clocks. Thus the clocks will be synchronized. A spaceship approaches the line of clocks at close to the speed of light. When the spaceship passes Station 1, its on-board clock and the master clock of the first station, Clock 1, read the same time, say, 03:00, by a previous arrangement. The watchman on Station 1 notes that the clock on the spaceship is running slow relative to his clock. By reciprocity, an officer on the spaceship will find that Clock 1 is running slow. The spaceship passes the line of space stations, reaching Station 5 at 03:10, according to the shipboard clock. What does the master clock of Station 5 read at this instant?

A clock puzzle

Begin by considering what the officer on board the spaceship sees. To him, all five clocks aboard the space stations are moving, and hence are running slow. Therefore, since the station clocks are all synchronized, Clock 5 must read the same as Clock 1, and as they are running slow, Clock 5 must read a time less than 03:10 when the ship passes. That sounds perfectly reasonable until we consider the point of view of the observers on the space stations. They see the spaceship rushing past them, and observe its on-board clock running slow in their frame. Since Clock 5 reads the same as Clock 1, Clock 5 must read a time later than 03:10 as the ship passes. This too sounds fine, until we realize that it is exactly the opposite conclusion to that which we reached by reasoning from the point of view of the officer on the spaceship. What is wrong?

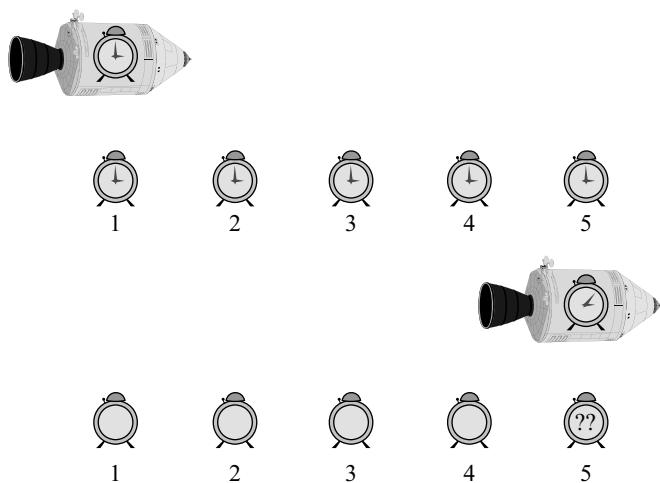


Fig. 7.14 A clock paradox. When the spaceship passes the first clock, all the station-based clocks are synchronized in the space frame. The spaceship's clock is in agreement with clock 1 as it passes. What does clock 5 read when the spaceship passes it at a later time? Is it ahead of or behind that of the spaceship?

Clock synchronization is a relative concept

The answer is that not all the statements made about the spaceship's point of view can be correct. We asserted in both arguments that the clocks were synchronized, but that is a statement about simultaneity which, we have learned, is relative. The station clocks are synchronized in their own rest frame, *not* in the frame of the ship. The five space-time events that correspond to the moment when the station clocks read 03:00 do not occur simultaneously in the frame of the ship. In the ship's frame, Clock 5 reads 03:00 at an earlier time than Clock 4 reads that time, which in turn reads 03:00 earlier than does Clock 3, etc. As the spaceship passes each clock, it sees it running slow, but they were never synchronized in its frame. In the ship's frame, Clock 5 was ahead of Clock 1, so when the spaceship passes this last station, Clock 5 is still ahead of the spaceship's clock, and therefore it reads a time later than 03:10. This is the careful explanation. The second argument above reached the correct conclusion because it referred only to observations made in the frame in which the station clocks were, in fact, synchronized. The first argument came to an incorrect conclusion because it attempted to make use of a synchronization that did not exist in the frame where it was employed.

Some misconceptions

Relativistic effects are not due to the finite travel time of light

It is sometimes believed that clocks seem to run slow due to the fact that the light from a receding clock takes time to catch up with the observer, so that time dilation is just an effect of the travel time of light. But by that reasoning, should a clock not run fast as it approaches? It is true that what we actually see, in terms of light striking our eyes, will be influenced by light-propagation effects, but this is a separate issue. It is important to realize that relativistic effects such as time dilation and length contraction do *not* arise because of any failure to take into account the finite travel time of the light. All observers know the speed

of light and are able to make use of this information in computing what the clocks in other frames are reading. For example, if in the clock paradox the spaceship officer continued to watch Clock 1, which might be, for example, a radio signal, while passing the other space stations, he would see Clock 1 running slow as it receded. But this has nothing to do with the fact that the radio waves must travel farther and farther to reach the ship. The on-board observer has a clock and knows how fast Clock 1 is moving relative to the ship, and can take into account the travel time of light when he computes the reading of Clock 1. The clock still runs slow relative to the shipboard clock. Similarly, when we synchronized the clocks in the space stations' inertial reference frame, we implicitly assumed that the speed of light was accommodated in our startup apparatus.

It is also widely believed that it might be possible to find a way to travel faster than the speed of light if only we would try hard enough, that scientists who say it cannot be done are fogies who cannot imagine new technologies. There are many reasons why it might be asserted that something cannot be done. One reason is simple ignorance. An example of this is the famous editorial of 1920 in the *New York Times*, which claimed that Robert Goddard's rockets could not possibly operate in a vacuum since, according to the writer, "Professor Goddard ... does not know the relation of action to reaction, and of the need to have something better than a vacuum against which to react—to say that would be absurd." We have studied Newton's laws of motion, and the flaw in this argument should be obvious. Of course, the reaction is against the *rocket*, or the jet airplane, for that matter, not against the air.¹⁰

Another reason that something cannot be done is inadequate technology. For example, until the 1940s many people believed that the speed of sound could not be exceeded. They based this belief not on any fundamental physical principles, but on the fear that the large stresses induced by supersonic travel could not be withstood by any material. It is correct that supersonic speeds create severe stresses and heating; supersonic aircraft must be specially designed, and are built with unusual materials. (The fastest acknowledged aircraft in the world, the SR-71 Blackbird, has a hull of titanium, an exceptionally strong metal.) But successful designs were created, and until recently it was even possible for the everyday wealthy and famous, not just military fighter pilots, to break the sound barrier in supersonic commercial jetliners. It was always recognized, however, that the sound barrier was an engineering problem, not a physical limitation. The speed of light is intrinsically different; it is not a mere technological challenge. It is a *fundamental* part of the way the universe works.

The speed of light is the ultimate speed limit.

¹⁰ Jet airplanes do depend upon the presence of air for steering; a rocket ship cannot bank against a vacuum, despite what might be depicted in movies. But both jet aircraft and rockets obtain their propulsion from the reaction of their gas exhausts against them.

The absolute limit of the speed of light is a basic aspect of the special theory of relativity. Denying this barrier would repudiate one of the most successful theories of modern physics. The special theory of relativity has been confirmed experimentally to almost fantastic precision. We have accelerated elementary particles to near the speed of light in large accelerators, and the relativistic effects we have discussed have been observed and measured. Special relativity has been subjected to some of the most exacting experiments ever performed, and it has in every case been found to give an accurate description of the observations. Denying relativity would also deny the validity of Maxwell's equations, and they are amply confirmed with each television or radio broadcast.

Relativity describes the behavior of the universe

The theory of relativity has profoundly altered the way in which we view the universe. It has merged concepts previously thought to be unrelated; space and time become space-time. Matter and energy are united into mass-energy. In special relativity, electromagnetism becomes consistent with mechanics. Special relativity also shows that electric and magnetic fields are essentially the same phenomenon. There is no ether, and hence no special frame at rest. The first postulate of relativity constrains all theories of physics, since all natural laws must be the same in any inertial frame. On the level of philosophy, the special theory of relativity eliminates the last vestige of Earth, or humanity, as a privileged observer, since it denies the existence of *any* preferred inertial frame. If special relativity seems to defy common sense and intuition, it simply means that the universe is more than our limited human awareness perceives it to be.

Chapter Summary

Einstein's theory of special relativity is based on two postulates: (1) the relativity principle, which states that the laws of nature are the same in all inertial reference frames, and (2) the speed of light in a vacuum is the same in all inertial frames. The second postulate is necessary in order that the laws of electromagnetism will obey the first postulate. All the consequences of special relativity follow from these two simple statements. A series of thought experiments shows that the second postulate of relativity leads to the conclusion that space and time intervals are relative, not invariant. Moving clocks run slow relative to a clock at rest, and a moving meter stick is contracted compared to the meter stick measured at rest. In both cases the amount of the *time dilation* or *length contraction* is specified by the boost factor $\Gamma = 1/(1 - v^2/c^2)^{1/2}$.

Space-time is the unification of space and time. A point in space-time, with three spatial coordinates and one time coordinate, is called an *event*. A sequence of events makes

up a worldline, a path through space-time. The invariant measure of the separation between two events in space-time is the space-time interval. Space-time separations can be divided into three categories. If the separation in time, multiplied by the speed of light, between two events is greater than their spatial separation, then their space-time interval is positive and they are said to be *timelike* separated. If they are separated by a larger space interval, they are *spacelike separated*. If two events can be connected by a single light beam, they are *lightlike* separated. Two events that can be causally connected must be separated by a timelike or a lightlike interval. *Proper time* is the time measured by an observer in his own rest frame; it corresponds to the space-time interval along the observer's worldline. *Proper length* is the length of an object as measured in its own rest frame.

The strange world of special relativity poses several apparent paradoxes. Studying these paradoxes helps to

understand relativistic effects. Special relativity tells us that many things that we intuitively regard as invariant are actually relative; for example, length, time interval, and simultaneity depend upon the relative motion of two observers. Some invariant quantities are the space-time

interval, the proper time, the proper length, and the rest mass. Despite the unfamiliarity of the relativistic world, it describes the behavior of the universe and has been well tested experimentally.

Key Term Definitions

Lorentz transformation The transformation, valid for all relative velocities, which describes how to relate coordinates and observations in one inertial frame to those in another such frame.

length contraction An apparent contraction of the length of an object in motion relative to a given observer, caused by the Lorentz transformation from one frame to another.

relativity principle The postulate of the special theory of relativity which states that the laws of physics are the same in all inertial frames of reference.

event A point in four-dimensional space-time; a location in both space and time.

simultaneity The coincidence of the time coordinate of two events; the observation that two occurrences take place at the same time. Simultaneity is not invariant, but depends upon the reference frame of the observer.

time dilation An apparent decrease in the rate of the flow of time (for example, the ticking of a clock) in a frame moving relative to a given observer, determined by the Lorentz transformation from one frame to the other.

proper time The time interval between two events as measured in the rest frame in which those events occurred. Numerically equal to the invariant space-time interval.

boost factor The quantity $\Gamma = 1/\sqrt{1 - v^2/c^2}$ that relates measurements in two inertial frames according to special relativity.

principle of reciprocity The principle in special relativity that two inertial frames will observe exactly the same phenomena when each observes the other. For example, each will see lengths in the other frame to be contracted by the same amount.

proper length The length of an object measured in its own rest frame.

rest energy The energy corresponding to the rest mass according to $E = m_0 c^2$.

space-time The geometry that merges space and time coordinates.

Minkowskian space-time The geometrically flat, four-dimensional space-time appropriate to special relativity.

space-time diagram A depiction of space-time, usually including time and only one spatial dimension.

worldline The path of a particle in space-time.

space-time interval The invariant distance in space-time between two events, as specified by the metric equation.

timelike Of a space-time interval: capable of being connected by something that travels below the speed of light in vacuo. Worldlines of massive physical objects follow timelike paths through space-time.

spacelike Of a space-time interval: incapable of being connected by anything that travels at or below the speed of light in vacuo.

lightlike Of a space-time interval: capable of being traversed only by a massless particle such as a photon. A lightlike, or null, space-time interval is zero.

lightcone The surface representing all possible paths of light that could arrive at or depart from a particular event.

past Those events that could have influenced a given event.

future Those events that could be influenced by a given event.

elsewhere Those events that cannot be causally connected to a given event.

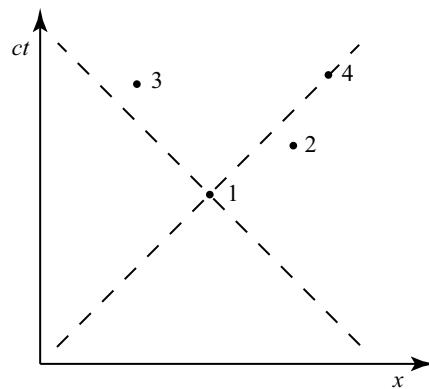
principle of causality The principle that a cause must always lie in the past of its effect for all possible observers.

Review Questions

- (7.1) How does the transition from Newtonian to relativistic mechanics illustrate one or more of the five criteria for a scientific theory?
- (7.2) A man watches a football game from the window of a train moving with constant velocity parallel to the football field. Consider the situation both from a Galilean and a Lorentzian point of view. Does the observer in the train measure the football field to be longer or shorter than the usual 100 yards? Is the length of the period as measured by the man in the train longer or shorter than the usual fifteen minutes?
- (7.3) If it is impossible to exceed the speed of light, why is it nevertheless possible to reach Alpha Centauri, a distance of 4 lightyears away, in less than 4 years of time, as measured by a space traveler?
- (7.4) A relativistic train approaches at a speed of $0.9c$. What is its boost factor? If the train has a length of 50 meters in its rest frame, what is its length in your frame? If you were the dispatcher of the relativistic train line, you would have to keep track of the clocks on the train as well as your own clock. If your station clock records an interval of 10,000 seconds between two given events, how many seconds does the conductor's on-board clock measure between the same two events?
- (7.5) In what way can time be called the fourth dimension? How does it differ from the other three?
- (7.6) Two spaceships (A and B) approach you at $9/10$ ths the speed of light ($0.9c$) from opposite directions. They send out radio messages. What is the speed you measure for the radio waves from A and B? What is the speed of the radio waves from A as measured by B? What speed does B measure for your motion? How fast does spaceship B observe spaceship A to be moving?
- (7.7) Consider the situation as depicted in Question 7.6. Whose clock (his, yours, or A's) does B observe to run slowest? Whose clock does A observe to run slowest? If all three ships have the same length when measured in their own rest frames, which ship does B observe to be shortest? If spaceship A sends out a pulse of red light, will B see it to be blueshifted or redshifted?
- (7.8) Draw a space-time diagram. Mark two events on it, labeled A and B. Indicate their separations in

space and time, that is, their Δx and Δt . Draw the worldline of a stationary observer. Draw the worldline of an observer moving at constant nonzero velocity. Draw the worldline of an accelerating observer. Draw another space-time diagram and include a moving inertial observer. Draw the space and time axes that correspond to the moving observer's rest frame and label these axes x' and ct' .

- (7.9) In the accompanying space-time diagram, which pairs of events may be causally connected? Which cannot be causally connected?



- (7.10) Using $E = mc^2$, we found that 1 kg of mass has the energy equivalent of 20 megatons. Estimate how many megatons of energy would be required to accelerate a spaceship with a mass of one million kg to a speed of $0.99c$. What does this suggest to you regarding the practicality of space travel at relativistic speeds?
- (7.11) Explain why we are unaware of the effects of special relativity in our everyday lives.
- (7.12) Describe two quantities considered invariant in Newtonian physics that are relative in special relativity. Describe two new quantities that are now known to be invariant.
- (7.13) (More challenging) Suppose that a train robber decides to stop a train inside a tunnel. The proper length of the train is 60 m, while the proper length of the tunnel is 50 m. The train is traveling at $4/5$ the speed of light. According to proper lengths, the train would not fit inside the tunnel. But the robber plans to use relativity to his advantage. The

length of the moving train in the rest frame of the tunnel, and of the robber, is 36 m. The robber computes this and decides to trap the train inside the tunnel, since, in his frame, the train should fit. From the point of view of the train's engineer, however, the *tunnel* is only 30 m long, just half the length of the train. The engineer knows that his 60-meter train will not fit completely into the tunnel. The robber thinks that the train will fit, whereas the engineer is sure it will not. But ei-

ther the train will fit or it will not; it cannot do both. Who is correct? *Hints:* Consider the following events: the locomotive enters the tunnel. The locomotive reaches the end of the tunnel. The caboose enters the tunnel. The caboose reaches the end of the tunnel. Which events are necessarily causally connected? Which are not? Draw a space-time diagram and label these four events. In which order do these four events occur in the robber's frame? In the train's frame?

This page intentionally left blank

The General Theory of Relativity

8

Tis like this gravity, which holds the
Universe together, & none knows
what it is.

Ralph Waldo Emerson

The need for a general theory

Einstein's conviction that the universe obeyed the relativity principle led him not only to revise Newton's mathematical equations of mechanics, but even more drastically, to discard the concept of absolute space and time. It proved unnecessary to alter Maxwell's equations of electromagnetism, since they already obeyed the correct invariance law; it was their lack of Galilean invariance that had motivated Einstein in the first place. Special relativity thus brought mechanics and electromagnetics into full consistency. But what about Newton's other great contribution, his law of universal gravitation? The special theory describes the relationship between measurements in inertial frames and has nothing to say about gravity.

In our discussion so far, we have been tacitly assuming that our laboratories on the Earth constitute inertial frames. But this can be true only in approximation; we constantly experience a gravitational force, yet an object experiencing a force is accelerated and hence cannot reside in an inertial frame. Moreover, there is no way to shield ourselves from the gravity of the Earth, or of any other massive object. Any object with mass will produce a gravitational force, in accordance with Newton's law. The universe is filled with masses, and the gravitational force extends indefinitely; in principle, there is no point in the universe where gravity's influence does not reach. Newton's law of gravitation also requires that the magnitude of the force depend on the square of the distance between the masses. Yet we have just learned that distance is not absolute, but relative; how, then, can we accommodate an inverse square law? Which distance should we use? Does the force vary according to the frame of reference? Finally, Newton's law implies that gravitational force is felt instantaneously at a distance. But we now know that this cannot be, for nothing can propagate faster than the

Key Terms:

- freefall
- Newtonian equivalence principle
- Einstein equivalence principle
- gravitational redshift
- Mach's principle
- flat geometry
- spherical geometry
- hyperbolic geometry
- metric equation
- metric coefficient
- geodesic
- tidal force
- Riemannian geometry
- gravitational radiation
- gravitational wave

The need for a relativistic theory of gravity

speed of light. Somehow time, and a finite propagation speed, must be incorporated into the gravitational force law. In the light of all these considerations, how can we incorporate gravity into the theory of relativity? We need a more *general* theory that will accommodate all frames, both inertial and noninertial, and that can describe the effects of gravity. This broader theory is Einstein's general theory of relativity. And just as special relativity had surprising, even astonishing, consequences, so we will find remarkable consequences of general relativity that will forever change our view of the cosmos.

General relativity may at first seem like an impressively exotic theory, but its underpinnings can be seen in the realm of the ordinary. In this Space Age, most of us have seen films taken aboard the Space Shuttle and its predecessors, enabling us to visualize experiences in which gravity seems to be absent. In particular, consider the Space Shuttle approaching a malfunctioning satellite 400 km above the Earth. A spacewalking astronaut attempts to snare the satellite, but the slightest touch sets it spinning. Finally, the 75 kg astronaut catches the satellite, which on the surface of the Earth would have many times the weight of the astronaut; yet in orbit, he handles it as if it were made of foam. Back inside the crew quarters, the crew members float about the compartment as if they were filled with helium. The television commentator says that this all occurs because of the weightlessness of outer space. But why is outer space weightless? Is there no gravity in outer space? No, although this is a common misconception.¹ Newton's law of universal gravitation states that the gravitational force is inversely proportional to the square of the distance from the center of one body to the center of the other. The radius of the Earth is approximately 6,500 km. Since 400 km is only about 6% of this number, the force of gravity at the altitude of the Space Shuttle still has fully 88.7% of its value at the surface of the Earth. If the Earth's gravity is still present, why are the astronauts weightless in space?

The answer lies in the realization that the concept of an inertial frame, and the effects of gravity, are intimately linked. In Chapter 6, we discussed the Rotor ride at the amusement park. As the cylinder spins faster and faster, the riders are pressed against the wall of the tube. When some particular angular speed is reached, the floor of the Rotor suddenly drops away, leaving the riders hanging on the cylinder's wall. Somehow the effective gravity experienced by the riders has changed. A similar, more serious, device is the centrifuge used in astronaut training. It consists of a small car attached by a metal arm to a central hub. A motor in the hub drives the arm and the car to rotate in a circle. If you could ride in this car, your body would feel as if it were very, very heavy. Your weight would seem to increase as the car spun faster around the central hub. Soon you would find that you could scarcely raise your arm

Why do astronauts become weightless?

Inertial forces are a consequence of an accelerated reference frame

¹Science fiction movies of the 1950s often showed objects beginning to float around a spaceship's cabin after the craft had "left the Earth's gravitational sphere." The genre was aptly named, since this is completely fictional science.

at all. What forces are acting here? Newton's first law requires that a force act in order for circular motion to occur; without such a force, an object will move in a straight line. That is why the car must be attached to the hub by a heavy metal arm; otherwise it would fly off in a straight line, with unhappy consequences for the rider. The real force experienced by the car is called *centripetal force*; it is the force that acts toward the central hub and causes the car to execute circular motion. In the rotating frame of reference, however, the rider experiences *centrifugal force*, which acts away from the center. Physicists often refer to an inertial force such as centrifugal as a *fictitious force*, because it is an artifact of the rotation of your frame of reference. But how can a fictitious force make you dizzy, or seem to increase your weight unbearably, or pin you to the side of a metal cylinder?

All these phenomena are connected to frames of reference. As we have already discussed, fictitious, or inertial, forces occur when an observer is in an accelerated, or noninertial, frame of reference. Nonaccelerated, inertial frames do not experience these forces. We have learned that the special theory of relativity relates observations made in inertial frames to one another, and because inertial frames are special, we call it the special theory. But what makes inertial frames special? We all know they are special by experience. When you change from one inertial frame to another, you feel an acceleration that has real, palpable physical consequences. (It is not the fall from a high building that would kill you, it is the sudden deceleration at the pavement!) We are not asking here how to treat acceleration mathematically or to account for its effects. Special relativity is perfectly capable of dealing with acceleration *per se*. However, special relativity presupposes the existence of inertial frames. It accepts Newton's first law as valid, and defines an inertial frame as one in which that law holds; that is, any free particle executes strictly uniform motion. We now need to know what determines an inertial frame of reference in the first place, and what creates the accelerations we feel when we are not in an inertial frame.

The equivalence principle

Like the special theory, the general theory is derived from only a few simple, powerful postulates. The first clue in our development of general relativity can be found in our contemplation of the weightlessness of astronauts. We may think of the phenomenon of weightlessness as some sort of antigravity effect, but what it really represents is a good inertial frame. When in orbit, the space shuttle is falling around the Earth in a state of **freefall**. The shuttle, the astronauts, their equipment, and the target satellite are all falling together. As Galileo demonstrated, all objects fall at the same rate in a gravitational field, regardless of mass. When a body is freely falling it is weightless, and hence in the state of freefall it *feels* as though gravity has been canceled. This simple idea

will be developed into one of the fundamental principles of the general theory of relativity.

From Newton, we learned that a force exists whenever a body is accelerated, and that the constant of proportionality between the force and the acceleration is the mass of the body, according to the equation $F = ma$. The m in this equation is the inertial mass, and it measures the resistance of an object to being accelerated. Even in orbit, in a state of weightlessness, it is still necessary to contend with an object's inertia; it remains difficult to push around massive objects. Newton also gave us the law of universal gravitation, which tells us that the gravitational force on a body is proportional to its mass. Experiment has shown that the inertial and gravitational masses are equivalent. But why *should* these two masses be the same? After all, the electric force is dependent on electrical charge, a quantity that is unrelated to the inertial mass of the charged object. Why should the gravitational force not depend on some special gravitational charge, which we might call the gravitational mass m_g , rather than on the inertial mass? There is no *a priori* reason why gravitational force should have any connection whatsoever with inertial forces. Yet experiment has clearly shown that the acceleration due to the force of gravity acting upon a body is independent of the mass of that body; this could be true only if inertial and gravitational mass were equal and the force is proportional to the mass. In the absence of any other forces, such as air resistance, all bodies fall with the same acceleration under the influence of gravity.

Inertial forces result from the acceleration of a reference frame. This results in inertial forces that produce the *same acceleration* on every observer in the frame. In other words, inertial forces are, like the gravitational force, proportional to mass. In practical terms, this means that gravitational and inertial forces produce effects that are indistinguishable. While standing on the floor of your kitchen, you drop something and it falls. This does not surprise you. Now imagine that you are traveling in a spaceship far from any source of gravity, between the distant stars. The spaceship's main rocket engine is engaged, and the spaceship is accelerating at one g , the same acceleration as that due to gravity on the surface of the Earth. If you drop something, it will fall against the direction of the acceleration of the spacecraft. Is that what you would have expected? Sitting in a chair in your living room can be just like sitting in a chair in a spaceship whose engine is operating.

Now suppose your spacecraft assumes an orbit around some planet. An *orbit* is a state of perpetual freefall around another body; no power is required to maintain it, provided that no energy is lost due to some deceleration, such as friction from the tenuous outer edge of the planet's atmosphere. The Moon orbits the Earth because it is accelerated toward the Earth, in accordance with Newton's law, but since the Earth is curved and finite and the Moon has some tangential motion, the Moon never approaches the Earth's surface. Thus the Moon is constantly falling. In an orbiting, freely falling frame, the inertial forces such as centrifugal force exactly cancel the gravitational force. This explains

Gravitational and inertial masses are equivalent

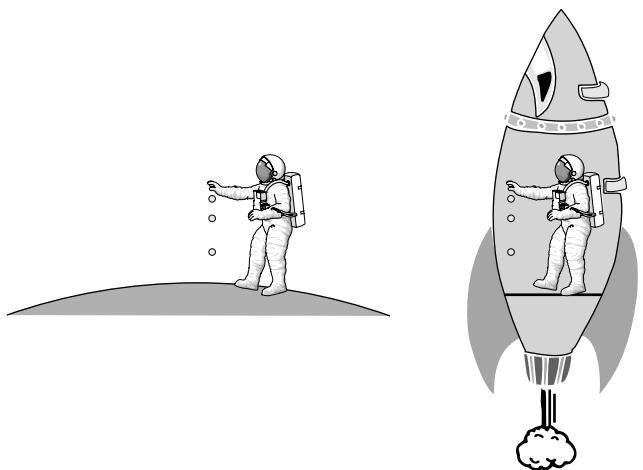


Fig. 8.1 The Newtonian equivalence principle states that gravity is indistinguishable from any other form of acceleration. The experience of a person in a spaceship accelerating with one g is the same as that of a person standing on a planet with gravitational acceleration g .

why the astronauts aboard the Shuttle are weightless while in orbit; it is *not* due to any lack of gravity in space. Gravity is canceled by freefall. This also explains the effect of the centrifugal forces felt in the Rotor ride or in the centrifuge. Those forces pull on you as if you were on a planet with a much stronger gravitational field. You cannot distinguish the inertial force from a gravitational force.

As part of their training, astronauts are exposed to weightlessness. Rather than drop them from a tower as in some modern amusement-park rides, NASA flies them in an airplane. For a portion of the flight, the airplane follows a parabolic trajectory that mimics the path of a freefalling object, such as a body thrown into the air. For a short but significant interval of time, the astronauts experience freefall and weightlessness. You may have had a similar experience if you have ever been aboard an airplane that encountered a severe stretch of turbulence and suddenly lost altitude. When the dinner trays and coffee cups start floating, that is a sign that the plane is momentarily in freefall.

The assertion that gravity is completely indistinguishable from (or equivalent to) any other acceleration is called the **Newtonian** (or weak) **equivalence principle**. It is not necessary to go into orbit to find examples of the Newtonian equivalence principle at work. For instance, suppose you enter an elevator to travel upward. The elevator starts and accelerates toward the top of the building, but you feel yourself pushed down against the floor. The force is the push of the floor upward on you, by which the elevator compels you to share its acceleration. In your frame, however, it is you that are pushing down on the floor, and this force is indistinguishable from a gravitational force toward the floor. You feel heavier than normal. When you return to ground level via the elevator, the elevator begins to accelerate downward. You feel lighter on your feet and experience a fluttery feeling in your stomach, as though your viscera are floating. Can you explain why you feel these sensations? As the elevator starts downward, its acceleration is toward the ground. Therefore, the elevator floor is falling away from your feet, and pushes

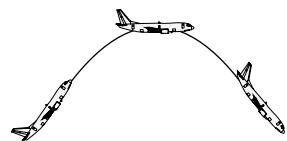


Fig. 8.2 A plane on a parabolic trajectory follows a freefalling path.

Can we distinguish an accelerating elevator from a changing gravitational field?

up on you with *less* force than when the elevator is stopped. Thus you press down on the floor with less force than you would if the elevator were at rest, and you experience this as a gravitational force that is less than the usual force due to the Earth. Notice that these effects occur only when the elevator is accelerating. After it has reached its constant operating speed, your weight will feel completely normal. But what if the cable were to break? You and the elevator would go into freefall down the shaft. The elevator and all its contents, including you, would then fall with acceleration g . Therefore, the floor of the elevator would exert *no* force upon you, nor would you exert a reaction force upon it. You would become weightless, not because gravity has suddenly been switched off, but because in your elevator frame, your apparent weight is the force you exert against the floor, and in freefall that force is zero.²

You might find it surprising that an acceleration *upward*, against gravity, makes you feel heavier. If gravity and acceleration are indistinguishable, then should not a downward acceleration feel like gravity? This counterintuitive effect stems from the fact that we all spend our lives in a noninertial frame, one tied to the surface of the Earth, and consequently we have difficulty in visualizing inertial frames. We are able to employ Newton's laws in our noninertial frame only by explicitly including the effects of gravity. When we drop a ball, we claim that it accelerates downward; that is, we are adopting the point of view that our Earth-based frame is inertial, in the sense of Newton's *second* law. We have now learned that our home frame is not an inertial frame, and would not be so even if it were to stop rotating, because it sits upon a large, gravitating mass. When we drop a ball, it occupies an inertial frame (temporarily, until it collides with the surface of the local gravity source) and is actually *not* accelerating while it falls. Thus it is *we* who are accelerated; and if a ball falling downward is not accelerated, then we must be accelerated upward. Hence what we call gravity is equivalent to an upward acceleration, as seen from the surface of the Earth. This is a subtle and perhaps difficult point, but important to a full understanding of the meaning of inertial frames in general relativity.

But what if you are sitting at rest in your living room? You will agree, no doubt, that you constantly experience gravity. But if you are sitting motionless, then how can you be accelerated? And if you are in free fall, are you not accelerated, with acceleration g , toward the center of the Earth? How can we reconcile our usual view of gravity as an acceleration with the claim that freely falling observers are unaccelerated? Perhaps the Rotor will again help to clarify these issues. To your friend watching you from an inertial frame, you are most certainly accelerated, else you would not be executing circular motion. Your friend watching from overhead says that you are experiencing a centripetal force, which is

²You cannot save yourself in a falling elevator by jumping up at the last moment. Since you would have a large downward velocity, jumping upward would decrease this by only a small amount. If your legs were strong enough to provide the required deceleration, they would also be strong enough to absorb the impact at the bottom of the shaft, as that too is just a deceleration.

provided by the wall of the metal cylinder. In your own frame, however, you are motionless; there is no centripetal force (toward the center), but rather a centrifugal force (away from the center), which exactly balances the reaction force from the wall. In your frame, you are not moving, so you claim that you are unaccelerated. There is only some force (gravity) pinning you to the wall of the Rotor. Gravity is just like the fictitious forces, for example, centrifugal and Coriolis forces, which we have previously identified as artifacts of a noninertial frame of reference. A freely falling observer is truly unaccelerated; it is you who are accelerating relative to the inertial frame. Yet just as we may still apply Newton's second law within a rotating frame, such as the Rotor or the Earth, provided that we introduce the fictitious forces to account for our noninertial motion, so may we introduce the fictitious force we call gravity, and continue to make use of Newton's second law for the conditions prevailing within our noninertial frame. From this argument it may seem as though we have just arrived at the conclusion that gravity does not exist! However, this is not the case. What it means is that gravity has no separate existence, but is related to the concepts of inertial frame and acceleration, which in turn are fundamentally tied to the nature of space and time. General relativity incorporates all these separate ideas into a unified picture.

Our discussion of the equivalence principle has so far been somewhat theoretical. Newton assumed the equivalence of gravitational and inertial mass, based on somewhat sketchy evidence and his intuitive sense of aesthetics. Yet this assertion can be tested. The earliest such experiments were performed long before the equivalence principle was formulated. Galileo discovered that bodies fall at a rate independent of their inertial mass, which motivated Newton to set the two forms of mass equal in the first place. Newton himself carried out experiments on pendula to test this hypothesis. He found no change in the period of pendula whose bobs were made of different substances, but were otherwise identical; of course, his experimental errors were large. The first highly accurate experiment to test the equivalence principle was performed in 1889 by Baron Roland von Eötvös. Eötvös constructed a device called a *torsion balance*. He suspended two bodies, of nearly equal mass but different composition, from a beam which hung from a very fine wire precisely at its center. If the magnitudes of the Coriolis force (from the Earth's rotation) and the gravitational force had differed between the bodies due to their differing composition, Eötvös would have been able to detect a twisting of the wire. None was seen, and Eötvös was able to conclude that inertial and gravitational mass were equal, to approximately one part in 10^9 . In the 20th century, Robert Dicke and others pushed the limit of such an experiment to the level of one part in 10^{11} , but the Baron's results were sufficient to convince many, including Einstein, that inertial and gravitational mass are equivalent.

To this point, we have confined our discussion to mechanics, the physics of motion. But the mechanical equivalence of inertial frames with freefalling frames hints at something deeper, namely the **Einstein**

Both inertial and gravitational force are proportional to mass

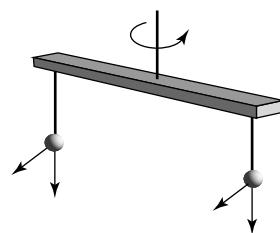


Fig. 8.3 A torsion balance can be used to test the equivalence principle. The spheres have identical mass, but are made of different substances; Baron Eötvös used wood and platinum. The spheres experience a gravitational force as well as an inertial force, namely, the Coriolis force due to the rotating Earth. If the ratio of the inertial to the gravitational mass in the two spheres was not exactly unity, a net twist on the wire would result.

The Eötvös experiment provides evidence for the equivalence principle

The laws of nature are invariant in all frames of reference, accelerated or not

(or strong) **equivalence principle**. This principle is similar to the first postulate of special relativity in its sweeping generality. The Einstein equivalence principle states that *all* inertial and freely falling frames are completely equivalent, and there is no experiment that can distinguish them. Whereas the Newtonian equivalence principle addresses only mechanics, the Einstein equivalence principle speaks of all of physics. In particular, it makes the relativity principle, and thus special relativity, applicable in freely falling laboratories. The Einstein equivalence principle is fundamental to general relativity; henceforth the expression *equivalence principle* shall refer only to the Einsteinian principle.

The equivalence principle has profound consequences, stemming from the fact that it is genuinely impossible to distinguish an inertial frame from a freefalling frame. A few thought experiments will clarify this. Imagine a beam of light, such as from a flashlight, shining from one side of an elevator to the opposite side. If this elevator were in a good inertial frame, perhaps in deep space far from any gravitational field, then an observer inside the elevator would see the beam trace a straight line across the elevator. Now consider the same situation in an elevator that is freely falling in a gravitational field. By the equivalence principle, we must observe exactly the same result as before: the beam passes straight across the elevator. But for the elevator in a gravitational field, this result tells us that the light must be falling along with all the other contents of the elevator. Otherwise, the elevator would fall some distance while the light was traversing it, and the beam would appear to the elevator-based observer to bend upward! This implies immediately that a gravitational field forces light to travel on a curved path, relative to the distant fixed stars. What does this mean for the theory of relativity? In the Minkowskian space-time of special relativity, light always travels on straight, lightlike lines through space-time. Now we find that the equivalence principle demands that these trajectories must curve in the presence of gravity. Hence the presence of gravitating matter affects space-time itself, changing inertial worldlines from the straight lines of special relativity to curved lines. The effects of gravity, then, can be incorporated into the theory of relativity by allowing space-time to curve. This is the fundamental basis of general relativity.

Such thought experiments can lead us to still more interesting results. Suppose a rocket in deep space accelerates forward. A bulb located in the nose of the rocket emits a beam of light, which is observed by a receiver on the rocket's floor. Because of the acceleration, the receiver's velocity will increase between the time of the emission of the light and its reception. This results in a relative motion between the receiver and the bulb at the moment of light emission, producing a blueshift in the light. According to the equivalence principle, the ship's forward acceleration is indistinguishable from a gravitational field directed toward its floor. The scenario is completely equivalent to an observer on the surface of the Earth receiving light from a bulb at a higher elevation. We have again shown something quite remarkable using only the equivalence principle: light traveling downward in a gravitational field, that

Light is affected by gravity

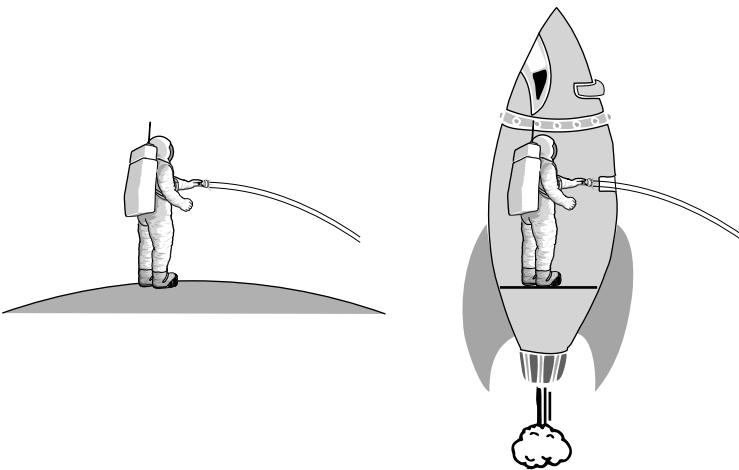


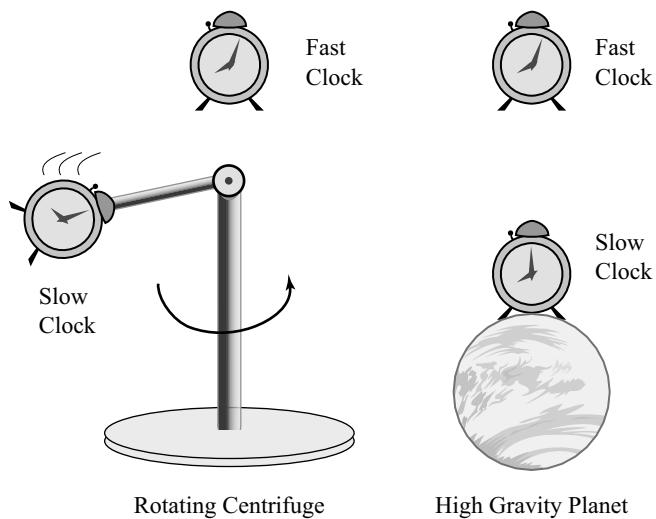
Fig. 8.4 The Einstein equivalence principle asserts that all laws of physics are identical in any inertial frame, including the laws governing the behavior of light. An observer in a spaceship accelerating with one g will see the light from the flashlight curve downward. An observer standing on a planet with a gravitational acceleration g will detect exactly the same curve in the light beam.

is, toward the source of the field, gains energy. An observer deep within a gravitational field will see a gravitational blueshift of light falling toward him. Conversely, a very similar argument demonstrates that light traveling upward, against a gravitational field, is redshifted, meaning that it loses energy. The loss of energy experienced by a light wave as it climbs out of a gravitational field is called the **gravitational redshift**.

Next we return to the centrifuge for another thought experiment. Imagine that you sit at the central hub, but have placed a clock on the end of the arm. You start the centrifuge and watch the clock spin around. As the centrifuge speeds up, the clock whirls faster and faster and experiences, in its own frame, greater and greater centrifugal acceleration. Since the clock is moving at a high rate of speed relative to your (nearly) inertial frame, by special relativity you will see the clock run slow. However, the clock is not in an inertial frame; it is accelerated. By the equivalence principle, this is the same as sitting in a gravitational field such as on the surface of a planet. From the equivalence principle, we thus conclude that a clock in a gravitational field runs slow relative to an observer in an outside inertial frame; thus we have discovered gravitational time dilation. Not only the clock, but all physical processes, run slow in the presence of gravity, so we cannot detect the effect by observing our own clocks. We must compare the ticking of clocks at two *different* points in a gravitational field. Unlike the situation in special relativity, there is no reciprocity for gravitational time dilation. Given two clocks carried by two observers, one deep in a gravitational field and the other high in orbit where the field is weaker, the clock at the lower point (stronger gravity) runs slower than the orbiting clock (weaker gravity), according to *both* observers. They disagree only in their judgment as to whose clock is running at the right speed, since both think their own is correct. We can use the equivalence principle to show that there is a purely gravitational length contraction as well, and similarly there is no reciprocity.

Gravitational time dilation and length contraction

Fig. 8.5 Time dilation from the equivalence principle. From special relativity, we know that the moving clock on the centrifuge will run slower than the clock at rest. The clock on the centrifuge is not in an inertial frame but is experiencing a centrifugal force. We conclude that inertial forces make clocks run slow. Gravity is equivalent to acceleration by the equivalence principle; hence a clock undergoing gravitational acceleration will run slow compared to an unaccelerated clock.



The GPS illustrates a real-world application of relativity

Real-world examples of general relativity are few and far between, but gravitational time dilation provides a few. Today's atomic clocks are sufficiently accurate that they can detect gravitational time dilation effects even here on the Earth. The atomic clock kept by the National Institute of Standards and Technology at Boulder, Colorado, at an altitude of approximately 1600 m, runs faster than the similar atomic clock near sea level, close to Washington, DC. The effect is small but detectable, and must be taken into account for high-precision measurements. Perhaps even more remarkable is the reliance of the Global Positioning System (GPS) upon relativistic corrections. GPS works by comparing time signals from an array of satellites, each carrying an atomic clock. Extreme precision of the timing information is critical to the performance of the system. An error of clock synchronization of 4 nanoseconds results in a positioning error of about a meter. As it turns out, relativistic effects are much larger than even this small tolerance, by a factor of nearly 10,000. There is a special-relativistic time dilation due to the satellites' motion with respect to the Earth-based observer, and a gravitational time dilation and blueshifting of the time signal due to the altitude of the satellites. If these relativistic effects were simply ignored, the accumulated position error over the course of 24 hours would amount to approximately 11 km, an error much greater than is permissible for the applications of the system.

Two viewpoints on the nature of space

What defines inertial frames?

The equivalence principle has led us to some interesting, even startling, conclusions about physics in a gravitational field. The equivalence between freefall and inertial motion, and between gravity and acceleration, suggests that somehow gravitating mass *defines* inertial motion

and acceleration, a significant departure from Newton's concepts. Let us examine two contrasting viewpoints regarding space and time.

In Newton's universe, acceleration is defined with respect to a space and time that exist absolutely, in their own right, independent of the existence of matter. They are simply laid down throughout the universe, and all action takes place with reference to them. They are unaltered and unaffected by any of the universe's events. Even special relativity, which blends space and time into space-time, does not question the existence of an absolute space-time. Accelerations occur with respect to absolute space or space-time. In the twin paradox, one twin feels an acceleration, the other does not, and this situation is not symmetric. Furthermore, absolute space postulates the existence of something that affects everything, but is itself affected by nothing. Gottfried Leibniz, Newton's contemporary and rival, first disputed this viewpoint, arguing that space cannot exist apart from matter. Admittedly, it is difficult to imagine what absolute space would be in the absence of matter, or how we would measure distances without some sort of yardstick. Although philosophers quarreled over this for two centuries, scientists overwhelmingly adopted Newton's viewpoint. Newton's mechanics *worked*, after all.

The first scientist to systematize the alternative worldview and place it into the context of physics, rather than metaphysics, was the physicist and philosopher Ernst Mach. The assertion that inertial frames are established only by the distribution of matter in the universe has come to be known as **Mach's principle**. Mach insisted that absolute space made no sense. To him, space, and hence inertial frames, are meaningful only in relation to the distribution of matter in the universe. Where there is nothing, one cannot define motion, much less acceleration. To illustrate this, consider an everyday example. Suspend a bucket of water from a rope, and give it a spin. Many of us have spun a bucket of water at one time or another, and we know that when the water starts to spin with the bucket, its surface curves, rising upward toward the rim of the bucket. Newton himself performed this experiment. The curving of the water is due to the centrifugal force, that is, to an inertial force resulting from the water's noninertial (rotating) frame. Now imagine a universe that is empty except for Newton's bucket of water, and repeat the experiment. If the universe contains nothing but this bucket of water, how does the water know that it is rotating? Rotating with respect to what? Would the water's surface still curve? Newton would answer that of course it would, since the bucket knows that it is rotating with respect to absolute space. Mach disagreed. To Mach, motion was inconceivable except in relation to other matter. The relative motions of matter determined acceleration. If there is nothing but the bucket, it cannot rotate with respect to anything, and rotation therefore has no meaning or significance.

Mach's principle states that the contents of the universe determine inertial motion

The Foucault pendulum provides another example. If we place a Foucault pendulum at the North Pole, it will swing in one plane with respect to the fixed stars while the Earth turns underneath it. Newton associ-

ated the realm of the fixed stars with absolute space; hence in the Newtonian view, the pendulum is moving inertially, with respect to absolute space. Now suppose that the fixed stars disappeared, leaving a universe consisting only of the Earth and the pendulum. How would the Foucault pendulum move then? According to Newton, the pendulum would be unaffected, because the fixed stars merely serve as convenient markers in absolute space. The Earth would continue to turn underneath the pendulum as it swung. Mach, however, would claim that the pendulum would now swing in a plane fixed on the Earth, since the Earth would be the only other matter present. In the absence of the remainder of the universe, there is nothing to define rotation for the Earth; hence inertial motion for the pendulum would be motion in a constant plane with respect to the only other matter in the universe, namely, the Earth.

It may seem hopelessly unrealistic to contemplate the consequences of thought experiments involving an empty universe, but if Mach is correct, there are observable effects even in our matter-filled universe. For example, the Foucault pendulum's motion, while dominated by the matter in the rest of the universe, must still show some influence from the nearby presence of the Earth. The rotating Earth must have *some* say in what constitutes a local inertial frame, so the local inertial frame must share in the Earth's rotation to some slight degree. This constitutes a kind of dragging of inertial frames, an influence that could, in principle, be measured. Thus, while the debate may have at first seemed to be about some esoteric, unresolvable issue, we see that there are real physical consequences and differences between the two points of view. Mach's ideas, particularly the suggestion that the overall distribution of matter in the universe determines local motion, heavily influenced Einstein's thinking. Einstein took the viewpoint that matter determines which trajectories will be freefalling, and hence inertial. The problem facing Einstein was to determine *how* matter could establish inertial frames. First, we must consider how we can define such frames in the presence of gravity; we have already shown that no frame that feels an influence of gravity can be inertial, and yet we know that gravity emanates somehow from matter. A more careful examination of the equivalence principle might provide a clue.

In adopting the equivalence principle, we have expanded the domain of special relativity. Previously, inertial motion was always straight-line motion, whereas it now involves curves, such as the curve of an orbit. We have also narrowed its scope, however, because we must now restrict ourselves to *local* measurements in small laboratories. Freefall is determined by the presence of gravitating masses, and the universe contains multiple overlapping and spatially varying gravitational fields. Inertial reference frames must be finite, therefore, because observers can be freefalling together only if the gravitational field they experience is uniform. For example, if one parachutist jumps from an airplane over the North Pole and another jumps over the South Pole, both are in inertial, freefalling frames, but those two frames are accelerating with respect to one other. Similarly, two skydivers falling side by side toward

Newton's and Mach's hypotheses make distinguishable predictions

the center of the Earth are moving inertially, yet they converge toward one another, even though they started off with what appeared to be completely parallel motion in one common inertial frame. These ideas hint at the basis for general relativity; matter exerts its influence through its effect on the *geometry* of space-time.

An introduction to geometry

Before we continue with our study of general relativity, we must take a detour through geometry. But what is geometry? It is the mathematics that describes the relationships of space, volumes, and areas. The typical secondary-school course on this subject seems to consist mainly of carrying out proofs of geometric propositions: the congruence of angles, similarity of triangles, and so forth. The major purpose is not so much to teach the applications of geometry as it is to teach the process of drawing logical deductions from a set of postulates. Like all mathematical and logical systems, geometry is built upon a set of obvious assertions, which we call postulates or axioms. All of the system is contained within the axioms. These postulates cannot be proven themselves, but have consequences that we can deduce. We have already seen an example of such a deductive system when we studied how the special theory of relativity was derived from two simple assertions.

The geometry of our high-school days is based upon a set of five postulates systematized by the Greek geometer Euclid. The resulting geometry is called, appropriately enough, Euclidean, and its postulates are as follows:

- (1) It is possible to draw a straight line from any given point to any other point.

The axioms of geometry

Note that we have defined neither “point” nor “straight line.” Do these concepts seem obvious? Like an axiom, the concept of a point is not definable within the Euclidean system, though we may define a straight line as the shortest distance between two given points. But now we have not defined what we mean by shortest. Does that also seem obvious? We shall find that it *does* require a definition, and we shall soon provide one, although Euclid himself, and his contemporaries, probably took it as another self-evident concept.

- (2) A straight line of finite length can be extended indefinitely, still in a straight line.
- (3) A circle can be described with any point as its center and any distance as its radius.
- (4) All right angles are equal.
- (5) Given a line and a point not on the line, only one line can be drawn through that point that will be parallel to the first line.

For centuries, mathematicians were suspicious of the fifth postulate. It seemed as though it should be a provable statement, not an axiom, and

some very distinguished mathematicians attempted to find proofs. All were flawed, but the struggle continued until the 19th century. The final acceptance that the fifth postulate is, indeed, a postulate, came from the independent demonstration by Carl Friedrich Gauss, Janos Bolyai, and Nicholai Lobachevsky that perfectly consistent geometries could be constructed if the fifth postulate were replaced by some other axiom. These geometries are said to be *non-Euclidean*.

If we accept the fifth postulate, then we can prove numerous geometrical theorems, of which two will serve as examples. These are that the interior angles of a triangle sum to 180 degrees, and that the circumference of a circle is equal to $2\pi R$, where R is the radius of the circle. Both these theorems are familiar to nearly everyone, and most of us take it for granted that they are facts. Yet they are valid only for Euclidean geometry. Euclidean geometry is **flat**; it is the geometry of a set of planes. Non-Euclidean geometries describe *curved* spaces. These spaces may have properties quite different from those of the flat Euclidean space. At first, these geometries may seem so strange as to be unimaginable; and to many people, what they cannot imagine must be impossible.³ But the non-Euclidean geometries are just as real as Euclidean geometry; it merely requires more reflection to think about them, since they are unfamiliar. To study general relativity, we must abandon our prejudices for flat space, and grant equal status to curved space.

It is easier to think about geometry if we start by considering only two-dimensional geometrical surfaces. An example of a curved geometry, one that should be easily imaginable, is the surface of a sphere. Where might we apply such a **spherical** geometry in our everyday life? To the surface of the Earth, of course, which is a sphere to a good enough approximation for most purposes. On a sphere, the equivalent of the straight line is the *great circle*, a circle whose center coincides with the center of the sphere. Slicing through a sphere along a great circle exactly bisects the sphere. On the Earth the lines of longitude are great circles, but with the exception of the equator, lines of latitude are not.⁴ Thus the circumference of a line of latitude depends on the location of the line. For example, you can walk along the entire length of a line of latitude by walking along a little circle centered on the North Pole. (Visitors to either of the poles sometimes do this and then claim that they have walked around the world. You could make a similar claim with nearly the same validity by performing a little pirouette in your backyard, since lines of latitude are arbitrary coordinates upon the sphere.)

The great circle is truly the equivalent of the straight line, in the sense that the shortest distance between any two given points on the surface of the Earth follows an arc of a great circle. This is why airplanes

Euclidean geometry is flat geometry

The surface of the Earth is a two-dimensional spherical geometry

³Conversely, many people are inclined to believe that whatever they can imagine must occur. Neither attitude is defensible.

⁴Lines of latitude are often called parallels, because when they are drawn on a flat map, they are parallel to the equator. But this is merely a conventional expression that has nothing to do with the geometry of the sphere.

fly along great-circle routes whenever possible. When flying from Los Angeles to London, aircraft do not fly directly east from Los Angeles; the route travels to the north, over Montana and Canada, across the North Atlantic and into London via Scotland. If you plot this route on a flat map, which is the projection of the surface of the sphere onto the plane, and is therefore always distorted, this does not seem like the shortest route at all. If the Earth really were flat, it would, indeed, not be the shortest distance. But if you plot the same route on a globe, you will easily see that it is the best one. Consider any two of these globe-girdling great circles, as shown in Figure 8.6. With a little thought, you will quickly realize that they intersect *twice*. There are no parallel lines, in the sense of Euclid's Fifth Postulate, on the surface of a sphere; all straight lines intersect twice. One way in which we can define the geometry of the sphere is by retaining the first four Euclidean postulates, and replacing the fifth with the statement:

- (5) Given a line and a point not on the line, NO line can be drawn through that point which is parallel to the first line.

Another property of spherical geometry is that the sum of the interior angles of a triangle drawn on the surface of the sphere is greater than 180 degrees. For example, consider a triangle made up of the portion of some line of longitude between the equator and the North Pole, another such line exactly 90 degrees from it at the pole, and the equator. The interior angles of the resulting triangle are all 90 degrees, for a sum of 270 degrees. Now pick a point on the surface of the sphere. Locate all the points that are an equal distance r from that point. By connecting those points you have drawn a circle, in accordance with Euclid's third postulate. But what is the circumference of that circle, and how is it related to the distance r from the center point? On a flat plane the circumference equals $2\pi r$, but on a sphere the circumference is *less than* $2\pi r$. This can be demonstrated on a familiar kind of sphere; on a rubber playground ball, imagine drawing a circle along points equidistant from the inflation valve, as measured along the surface of the ball. Cut along the circle. You now have a little cap of rubber. If you try to press it flat, thereby forcing the spherical-geometry circle onto a flat geometry, the rubber will stretch or tear; there is not enough material to make a flat circle. Finally, this spherical geometry is finite, yet has no edges. If you travel along any great circle (that is, along a straight line) long enough, you will end up precisely where you began. In Euclidean geometry, in contrast, straight lines have infinite length.

So far we have obtained two different geometries by assuming that either no, or one, line can be drawn parallel to a given line, through a given point. What if we assume that more than one parallel line can be drawn through such a point? This turns out to be the same as allowing an *infinite* number of parallel lines to be drawn through a point. Such a geometry may seem very strange, too strange to imagine, and indeed this **hyperbolic** geometry cannot be constructed, even in its two-dimensional form, in three-dimensional Euclidean space. It is

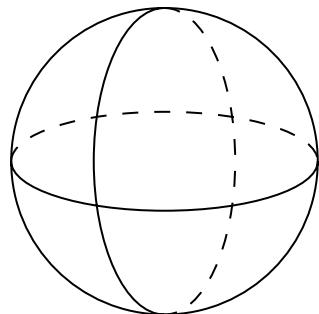


Fig. 8.6 Great circles upon a sphere. In spherical geometry, any two great circles intersect at exactly two points.

A new concept: hyperbolic geometry

extremely difficult to visualize this geometry. Yet it is as self-consistent as spherical or flat geometry. The properties of this geometry are in some respects exactly opposite to those of the spherical geometry just described; interior angles of a triangle sum to less than 180 degrees, there is an infinite number of parallel lines through a point, and the circumference of a circle is greater than $2\pi r$. This geometry is also, like Euclidean geometry, infinite, but in some sense it is still larger than a Euclidean space. In a three-dimensional hyperbolic space, there is more volume contained within a given radius than is contained in the corresponding radius within a Euclidean space. Although the hyperbolic geometry cannot be visualized, a saddle exhibits some of its properties near the *saddle point* at the center; it can be employed as an aid to the imagination. At the saddle point, the surface curves up in one direction and down in the other. Mentally draw, and cut out, a little circle of some small radius around the center of the saddle,⁵ such that the circle contains material going both uphill and downhill. If you try to crush this circle flat, you will find you have too much material; there is overlap. This shows that circles in hyperbolic geometry are larger than the corresponding circles in Euclidean geometry. The saddle, a two-dimensional surface embedded in a three-dimensional flat space, has this property only at the saddle point. The hyperbolic geometry exhibits this property at *every* point.

In discussing the spherical and hyperbolic geometries, we have used two-dimensional examples. Both these geometries have three-dimensional forms as well, just as Euclidean geometry has two-dimensional (planar) and three-dimensional versions. We used the example of the surface of the Earth, a two-dimensional spherical geometry, because it is familiar to everyone and because we can visualize it. If we extend the spherical geometry to three dimensions, it retains all the properties described above, with appropriate additions for the third spatial dimension, but we can no longer visualize it. (Mathematically, a three-dimensional sphere is *not* a three-dimensional ball. It is the surface of a four-dimensional ball.) Similarly, the hyperbolic space can be described in three dimensions, but since we cannot even adequately visualize a two-dimensional hyperbolic surface, we have no chance of imagining the appearance of the hyperbolic space in higher dimensions. Nevertheless, the mathematics is essentially the same, regardless of how many dimensions we use.

Most of us probably think of geometries in terms of two-dimensional surfaces that exist within the three-dimensional Euclidean space of our experience. This can be misleading. All geometries have some properties that are intrinsic; these properties do not depend upon any higher-dimensional entity. A sphere has an intrinsic curvature that is a property of the geometry itself, and does not depend upon the sphere existing within, or being embedded in, a three-dimensional Euclidean space. The surface of the Earth constitutes a two-dimensional sphere, to a good approximation. If we were two-dimensional creatures, we would be in-

Geometries come in arbitrary dimensions

⁵A *Pringles*™ brand potato chip provides a reasonable experimental model.

capable of visualizing the third dimension, so we would have no direct knowledge of the embedding of our sphere within any higher dimension; yet we could, by local measurements, determine that our geometry is curved. Furthermore, the existence of curvature in a geometry does *not* require a higher dimension in which the geometry curves. Mathematicians showed long ago that a geometry need not be embedded in anything at all, but can exist independently. In classical general relativity, the universe consists of a four-dimensional curved space-time geometry; it is not embedded in something of even higher dimension.

We have discussed three types of geometry: spherical, flat and hyperbolic. These three examples are special cases from a whole host of possible geometries. What makes these geometries special is that they have the same properties everywhere, that is, they are all *homogeneous*; and they have no special directions, that is, they are all *isotropic*. One point in Euclidean flat space is the same as any other. One point on a sphere is like any other. (On the Earth we regard the North and South Poles as special locations because the Earth is rotating, and the poles lie along the axis of rotation; this is a physical, not a geometrical, property.) Similarly, the hyperbolic geometry is the same at all locations. General relativity is not restricted to these specific geometries, but they have a special role to play in cosmology, because they are possible geometries for a homogeneous and isotropic universe. For now, however, we must discuss a few formal mathematical considerations of generalized geometries, as a way of understanding how general relativity works.

The metric equation

Suppose we wish to measure the distance between two points in one of these generalized geometries. We already know how to do that in Euclidean space; on the plane, the distance is given by the Pythagorean Theorem as

$$\Delta r^2 = \Delta x^2 + \Delta y^2, \quad (8.1)$$

where Δx and Δy are the distances given in perpendicular coordinates (x, y) , laid out like a grid on our flat plane. For general geometries, we must write the analogous formula as

$$\Delta r^2 = f \Delta x^2 + 2g \Delta x \Delta y + h \Delta y^2, \quad (8.2)$$

where Δx and Δy are still our grid of coordinates, but now the grid follows the folds and curves of the geometry. This requires that we introduce the additional functions f , g , and h , which depend on the geometry. This formula is called the **metric equation**, and the quantities f , g , and h are the **metric coefficients**. We have expressed the distance in these formulas as Δr in order to emphasize that it represents the distance between two points, and thus can be taken as a small increment itself, as signified by the Greek letter Δ .

Since the metric coefficients depend upon the geometry, the general metric equation written here gives the distance between two points only

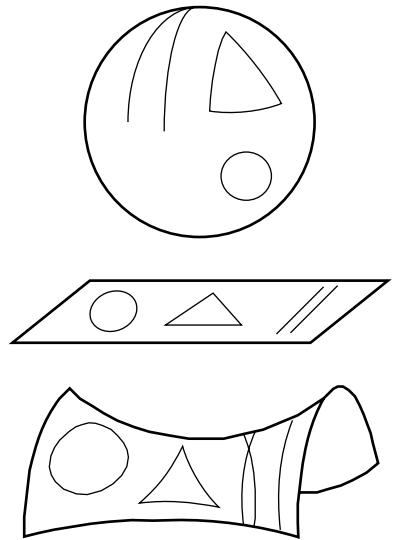


Fig. 8.7 The three types of isotropic, homogeneous geometries and some of their properties. From top to bottom, these are the spherical geometry, the flat geometry, and the hyperbolic geometry.

The cosmological principle limits the possible geometries of the universe

A prescription for distance in an arbitrary geometry

in the case that those points are very close together, so that the values of the coefficients change little from one point to the next. Ideally, the points become so close together that their separation is infinitesimal. In order to compute the distance between two arbitrary points, we must know not only their coordinates and the metric equation, but also the path along which we wish to find the distance. This should not be surprising, since we all know from our everyday experience that the distance between two points depends upon the path taken from one to the other. We compute the total distance by summing the small incremental distances along the path or, in the language of calculus, by integrating along the path. The metric equation is useful for more than just the distance between two points, however; it can also be used to calculate more complicated quantities that specify the curvature intrinsic to the geometry. The metric equation and the coordinates together describe the properties of the geometry.

Most geometries do not have a metric as simple as equation (8.1). For example, on the spherical surface of the Earth, we almost always use latitude and longitude as coordinates. To find the distance between two points on the Earth, we must use a more complicated metric equation given by

$$\Delta r^2 = R^2 \Delta\theta^2 + R^2 \cos^2 \theta \Delta\phi^2, \quad (8.3)$$

where R is the radius of the sphere, θ is the latitude from the equator of the sphere, and ϕ is the longitude. In this example, the metric coefficients are $f = R^2$, $g = 0$, and $h = R^2 \cos^2 \theta$. The h metric coefficient tells us, for example, that traveling 20° to the west in longitude is considerably farther at the equator, where $\theta = 0$ and $\cos \theta = 1$, than it is near the North Pole, where $\theta = 90^\circ$ and $\cos \theta = 0$.

The distance along a given path between two points is a real physical property relating those points, and it does *not* depend upon the coordinates; the role of the metric equation is to describe how to compute this distance, given a particular set of coordinates. Some paths connecting the points may be special in some regard. For example, in Euclidean space there exists a unique path between any two points whose distance is the shortest possible; it is the straight line that connects the two points according to Pythagoras' theorem. For two-dimensional Euclidean space, the length of this straight line is given immediately by equation (8.1), provided that we use the so-called Cartesian coordinates (x, y) that are everywhere perpendicular to one another. Cartesian coordinates happen to be the right coordinates for Euclidean space, in the sense that their corresponding metric equation is the simplest possible, but other coordinate systems may be used, with appropriate changes to the metric equation. If we choose to employ some coordinate system other than the usual (x, y) system in our ordinary Euclidean space, the metric coefficients will vary from one point to another, and computing distances is more awkward; regardless, the distance does not depend upon the choice of coordinates. When we extend our concept of distance to more general spaces and coordinates, the metric coefficients

A distance along a path is independent of the coordinates used to measure it

will usually depend upon the location for all choices of coordinates, due to intrinsic curvatures in the geometry.

The space-time interval is the distance between two events, and it is also an invariant physical property, independent of the coordinates Δt and Δx . Since the metric specifies distances in the geometry, there must exist a general metric for space-time, which we can write in the form

$$\Delta s^2 = \alpha c^2 \Delta t^2 - \beta c \Delta t \Delta x - \gamma \Delta x^2, \quad (8.4)$$

A general form for the space-time interval

where, for simplicity, we have expressed only one of the spatial coordinates, x . In the present discussion we have been restricting ourselves to spatial relations only, but we shall soon make use of the space-time interval in curved space-times and general coordinates.

The structure of general relativity

We now have in place all the parts we need to complete the description of general relativity. We have learned that the equivalence principle implies that masses define inertial trajectories. We have seen how it is possible to construct geometries other than our usual flat Euclidean geometry, and thence to use geometry to define the equivalents of straight and parallel lines. Now we must complete the task by showing how mass determines geometry, and geometry determines inertial trajectories.

To begin, recall that special relativity showed us how to relate observations made in one inertial frame to those made in any other inertial frame. We found that there is an invariant quantity, the space-time interval. The space-time interval between two events along any particular worldline, which need not be inertial, corresponds to the proper time measured by a clock traveling on that worldline. All observers will agree about this proper time, although they may not agree about the rate of ticking of the clock. In special relativity, our inertial observers move at constant velocity along a straight line through both space and space-time. How can we define an inertial observer in the more general case of a curved space-time? We can do so by recalling that in special relativity, the proper time interval between two events always has its greatest possible value along the worldline of an inertial observer. In going to curved space-times, we must generalize this idea somewhat. Any path between two distinct points that is an *extremum*, that is, the longest or shortest possible, is called a **geodesic**. In special relativity, the geodesic is a straight line through space-time, and always has the *maximal* value of the space-time interval or proper time.⁶ We can immediately generalize this concept to curved space-times. Any observer traveling along a geodesic in space-time is an unaccelerated, inertial observer. In general

The maximum proper time elapses along inertial paths through space-time

⁶The straight line is the shortest distance in a Euclidean space, but in Minkowskian space-time, because of the negative sign in the definition of the space-time interval, a straight line defines the *longest* proper time between two events. This is another way in which our intuition can be tripped up by special and general relativity.

space-times, these geodesic worldlines need not be the straight lines of the Minkowskian geometry; they are determined by the curvature of the space-time geometry.

How do we determine the geometry of space-time, so that we may compute the geodesics and find our inertial frames? This is where the equivalence principle shows the way. The equivalence principle states that inertial frames are identified with freely falling frames; they are completely equivalent. We know from experiment that freefall is determined by the distribution of masses producing what we have called gravity. Mass determines gravity, and gravity defines the inertial reference frames, or the geodesics, of the space-time. Thus mass by its very presence causes space-time itself to curve. Einstein's great contribution was to work out how this is accomplished through the geometry of space-time. How might this happen? We can construct a model that helps to visualize the idea. Imagine a sheet of rubber stretched flat, suspended between supports. The geometry on this sheet will be, of course, Euclidean. Draw straight, perpendicular coordinate lines upon the rubber surface. Now imagine that you scatter heavy steel ball bearings across the sheet. As you would expect, the bearings will distort the surface of the rubber. Lines that were straight (the geodesics) on the flat sheet now become curved. On the surface of the sheet, some of those geodesics will twist around the spheres (the orbits), while others will be deflected by the spheres; far away from any ball bearings, where the rubber is still almost flat, the geodesics will once again become straight lines. The ball bearings determine the geometry of the rubber sheet, even at quite a distance from them.

In our model, the rubber sheet is filled with ball bearings of various sizes, causing the geodesics to curve in complicated and elaborate ways. Particles moving without friction across such a surface naturally follow geodesics along that surface. The curvature determines the inertial (geodesic) motion at any point. Thus heavy ball bearings (matter) determine inertial motion of free particles through their effect on the geometry of the rubber sheet. In general relativity, masses alter the geometry of four-dimensional space-time, causing geodesics to be curved paths. The idea that masses determine inertial motion is similar to Mach's principle, although Mach never developed any formal way for matter to accomplish this task. Einstein himself intended to incorporate Mach's principle into the general theory of relativity, but it is present more in its spirit than in practice. In many cases, the equivalence principle makes different predictions from Mach's principle. The mean distribution of matter in the universe does establish the geodesics, but within an inertial frame of reference, the laws of physics feel no effect of any matter whatsoever. General relativity walks a middle ground between Newton and Mach. We cannot say for sure which picture is absolutely correct for our universe, but experimental evidence supports the equivalence principle.

Once Einstein had decided that the geometry of space-time fixes the inertial frames of reference, he had to establish the specific mathematical

The rubber sheet as an analogy to curved space-time

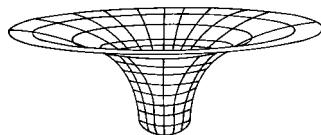


Fig. 8.8 A small portion of a rubber sheet geometry, showing its distortion by a massive ball bearing. Away from the ball, the sheet increasingly reverts to its flat state. Near the ball, the sheet is affected by the presence of the mass. A small test particle rolling on the sheet would be attracted to the ball due to the distortion of the rubber sheet.

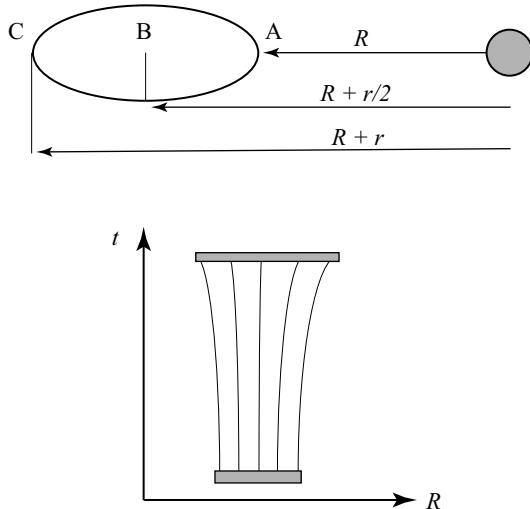
connection between geometry and gravity. This was the most difficult aspect, and it took him several years to find the right way. We will not trace the full development of mathematical general relativity, but we can outline some of the steps Einstein took. Let us begin by returning to the Minkowskian space-time of special relativity, which consists of a three-dimensional Euclidean space along with one time dimension. We know how to compute geodesics and find inertial frames in the Minkowskian space-time. Special relativistic inertial frames extend to spatial (and temporal) infinity. At each event, lightcones can be constructed that divide space-time into timelike, spacelike, and lightlike regions.

However, Minkowskian space-time is not adequate to describe space-time in the presence of gravity. The equivalence principle tells us that inertial frames in curved space-time must be freely falling. What else can we say about such inertial frames in the presence of gravity? We can easily deduce that they must be restricted in their extent by considering **tidal forces**. We are all familiar with tides, but what do we mean by a tidal force? Such a force occurs due to differences in the gravity acting over an extended body. For example, the feet of a human body are closer to the center of the Earth than is the head. By Newton's law of gravity, the gravitational force is not constant, but decreases over the distance from feet to head. Therefore, the feet experience a stronger attraction toward the center of the Earth than does the head. The difference in the attraction goes approximately as the ratio of the individual's height to the cube of the radius of the Earth. Luckily, this ratio is quite small, and for all we care, we live in a constant gravitational field. If the ratio were not small, we could be torn apart by the tidal force. The more rapidly the gravitational field changes over a given distance, the greater the tidal force. If an extended body in a gravitational field is to hold itself together in the face of strong tidal forces, some internal cohesion must be present. In most cases, at least in our solar system, intermolecular or self-gravity forces are adequate to keep the body intact. However, some comets have been observed to break apart due to the tidal forces encountered in passing close to the Sun, or to one of the major planets such as Jupiter.

Tidal force

Tidal forces do ultimately cause oceanic tides on Earth; hence their name. The Moon can be on only one side of the Earth at a time, so its pull on the Earth differs significantly across the diameter of the Earth, as illustrated by Figure 8.9(a). This causes the Earth to bulge on *both* sides. The Earth is mostly solid and has limited ability to change its shape in reaction to the changing and differential attraction of the Moon; but the waters of the oceans can, and do, flow in response. Therefore, tides occur roughly on opposite sides of the Earth at the same time. The exact behavior of water tides depends in quite complicated ways upon the topography and shapes of the basins in which the oceans are contained, and so for the seas this picture is greatly oversimplified, but it describes the basic driving force. The Sun also creates tidal forces, but although the Sun is many, many times more massive than the Moon, it is also much farther away, and tidal forces diminish even more rapidly with

Fig. 8.9 Tidal forces from two points of view. In the Newtonian view, the gravitational force on a planet due to a mass is GMm/R^2 at point A, $GMm/(R + [r/2])^2$ at point B, and $GMm/(R + r)^2$ at point C. The force at A is greater than the force at B, which in turn is greater than the force at C, and the planet is pulled into an ellipsoidal shape. In the relativistic view, tidal forces result from the shape of space-time. The lines are geodesic paths through space-time that diverge due to curved space-time. The diverging geodesics pull apart an extended body; the greater the divergence over a specified distance, the greater the tidal force.



distance (like $1/R^3$) than does the gravitational force. Consequently, the Sun's influence upon tides is less than half that of the Moon.

Tidal forces are an intrinsic property of gravitational fields. By the equivalence principle, a freely falling elevator above the Earth would be a local inertial frame, were it not for the presence of tidal forces. Since everything in the elevator is falling toward the center of the Earth, a ball on one side of the cabin and another on the other side will accelerate toward one other in the elevator frame, as their freefalling trajectories converge. In freefall, both balls follow geodesics in the region of space-time they occupy. Yet these geodesics are not parallel, because the balls approach one another as both fall toward the center of the Earth. From the point of view of geodesics, tidal forces result from the fact that geodesics in a curved geometry need not remain at some fixed separation. An extended body cannot travel on a single geodesic, and if the nearby geodesics upon which various parts of the body travel should diverge, then stresses will result that would tend to pull the object apart. Conversely, converging geodesics could compress an extended body.

Do tidal forces invalidate the equivalence principle? They might seem to provide us with a mechanism for distinguishing between an elevator falling toward the Earth and another floating in deep space. Actually, the equivalence principle remains valid, but the existence of tidal forces means that any inertial frame we might hope to construct must be small, in the sense that the tidal forces within it must be zero or very small. Thus inertial frames in general relativity are *local*; that is, they are valid only in the immediate vicinity of the freely falling observer. A single inertial reference frame cannot be defined to cover all space and time when gravitating masses are present. Hence space-time, in general, cannot be the special-relativistic Minkowskian space-time. However, within restricted, freefalling inertial frames, we know from the equivalence principle that the local geometry of space-time must be Minkowskian, that

is, flat. No matter what the overall geometry of space-time may be, for sufficiently small regions it must reduce locally to a flat space-time. That is, regardless of how space-time may curve, it must be possible to consider a region small enough that the curvature can be ignored. The surface of the Earth behaves analogously; although the Earth is spherical, and hence has a curved surface, it appears flat when observed locally, such as in a Kansas wheat field.

Within the realm of all conceivable geometries, only a very few special geometries have this property of local flatness. Mathematicians tell us that the most general geometries that are locally flat are those studied by Georg F. B. Riemann. Such geometries are called **Riemannian geometries**; they are characterized by invariant distances (for example, the space-time interval) that depend at most on the squares of the coordinate distances (Δx or Δt). For small enough regions, the metric equation must reduce to the familiar Pythagorean rule in space, or to the Minkowskian space-time interval in a relativistic space-time. Thus, *all possible* space-time geometries can be represented by the form given schematically in equation (8.4). This is a powerful notion, because it eliminates most of the infinity of possible geometries and restricts our candidates to Riemannian geometries. Matter creates some geometry that may be very complicated, but it cannot be arbitrary; it must be of Riemannian type. The mathematicians had already worked out a full set of equations describing these geometries. All we must do now is write an equation that connects gravity, and the mass that produces it, to geometry.

One hint of the form of the equation is the correspondence between tidal forces and certain mathematical properties of a Riemannian geometry. Any deviations from flatness in Riemannian geometries are specified by expressions for the curvature. Similarly, tidal forces describe the gravitational deviations from flatness in a local frame. Thus, the geometrical curvature must provide a mathematical measure of the physical tidal forces in a gravitational field. Another important clue that guided Einstein in his search was the requirement that his new law of gravity reduce to the Newtonian law for those cases in which velocities are much less than the speed of light and gravitational forces are weak. We are quite sure that Newton's law of gravity works very well for describing the trajectories of the planets and of spacecraft, so any new theory would have to be consistent with the old law under appropriate conditions. Although Einstein was developing a radically new theory, he was still stringently constrained by the success of Newton.

As we know, Einstein did succeed in deriving equations of general relativity that satisfied these requirements. The first complete publication of the general theory was a 1916 paper in *Annalen der Physik*, “Die Grundlage der allgemeinen Relativitätstheorie” (The foundation of the general theory of relativity), though portions of the equations had appeared earlier. In their most compact form, those equations are

$$G^{\mu\nu} = \frac{8\pi G}{c^4} T^{\mu\nu}. \quad (8.5)$$

Space-time is locally flat

General relativity is consistent with Newtonian gravity

Einstein's equations

No doubt this seems quite remarkable, to write down the entire universe in one line, but what do these symbols mean? The notation we are using here is very compact and comes from a branch of mathematics called *tensor analysis*; in fact, this single line represents ten complicated equations. But we can gain some insight without delving into the mathematics. The term on the left, $G^{\mu\nu}$, comes from mathematical geometry and describes the curvature properties of a four-dimensional Riemannian geometry. It consists of ten different components; the μ and ν are not exponents, but are labels for the various space and time components of the geometry term. The term $T^{\mu\nu}$ on the right-hand side has corresponding components in space and time; it is called the *stress-energy tensor*, and it contains the description of the matter and energy densities, pressures, stresses, and so forth, with which space-time is filled. The constant factors on the right are required for consistency of the units; our old friend, the gravitational constant G , appears in a prominent role even in general relativity.

The Einstein equations are very difficult to solve. With most other sets of equations, it is possible to start with the set of rules associated with the geometry, and coordinates can be chosen for convenience. Here those things are part of the solution. Einstein's equations are also highly nonlinear. This means that if solution A is found, which could be, for example, the gravity around a spherical star, then the gravity around two spherical stars is *not* A + A. Solutions do not simply add together, that is, superpose; the solution of a full system is more than the sum of its individual pieces. Consequently, Einstein's equations have been solved exactly only for a few simple cases.

The mathematics may be complicated, but in their essence these equations state that *Geometry = Matter + Energy*. Thus if matter or energy exists, it acts as a *source* for the geometry. This is not all that different from Newtonian gravitation, with the notable exception that now energy, in any form, is also a source of gravity. (Special relativity has already taught us that mass is just a form of energy, so we should have expected a result such as this.) But let us go further and suppose that no matter or energy is present, so that we are left with $G^{\mu\nu} = 0$; this is still a valid equation. Geometry exists regardless of the presence of matter. Gravity itself turns out to be a form of energy, so not only does matter create gravity, that is, curvature, but gravity acts back on itself to create gravity. Space-time curvature can exist, and even act dynamically, without the presence of any matter or nongravitational energy. This is one of the ways in which Einstein and Mach part company.

Einstein's equations finally overcame one of the problems with Newton's law of gravitation, by incorporating time and a finite propagation speed into gravity. This leads to a surprising consequence of the Einstein equations: moving masses can generate waves of curvature, or **gravitational radiation**. If the matter side of Einstein's equations changes, then the geometry will change as well. Thus a gravitational field that varies in time can produce a wave in the curvature of space-time itself; this is a **gravitational wave**. Gravitational radiation propagates away

Gravity, like mass, is a form of energy

The phenomenon of gravitational radiation

from its source at the speed of light. To return to the analogy of the rubber sheet, imagine that we allow the ball bearings to roll, sending small ripples throughout the sheet. Gravitational waves are ripples in space-time, like water waves that disturb the surface of a pond.

Tests of general relativity

It is a wonderful thing to have a beautiful theory of gravitation. But Einstein's equation (8.5) is not a proof; it is a hypothesis. In science, theories must be tested. The general theory of relativity satisfies our requirements for a scientific theory, since it makes many predictions that are useful for testing. However, all the gravitational fields that we have handy in the neighborhood of the Earth are extremely weak, in the sense that the curvature of space-time in our vicinity is not very far from flat. The differences between Newtonian and Einsteinian gravity are most pronounced for strong fields, such as that of the black hole, the topic of the next chapter. Unfortunately, or perhaps fortunately, no sources of very strong, and thus especially interesting, gravitational fields are immediately available to us for direct measurements. The weakness of our local fields complicates our efforts to test general relativity, since deviations from the predictions of Newtonian theory are small, and since agreement in the weak-field limit may not automatically extrapolate into the realms of strong fields. Even so, some ingenious experimental tests have been performed for the general theory.

One of the first predictions put to the test is the one mentioned early in our discussion of the equivalence principle, namely the bending of a beam of light in a gravitational field. We did not state so explicitly, but this would be predicted even within Newtonian theory, since we used nothing but the equivalence principle to obtain this result. However, general relativity goes further by recognizing that space-time near a massive object is itself curved, not flat. Hence the bending of starlight as it passes close to the Sun is greater than would be predicted from Newtonian gravitation, since the light both falls in the gravitational field and travels through a curved space-time. Specifically, the total bending of a light ray around the Sun would be twice as great in general relativity as it would be in a flat space-time with Newtonian gravitation, making this phenomenon an effective discriminator between the theories. Normally, one cannot see stars close in the sky to the Sun since they are lost in the Sun's glare, but during a total eclipse such stars become visible. It is then possible to make a careful determination of the location of the image of those stars during the eclipse. Those apparent positions can then be compared to the positions in the sky of the same stars during the part of the year when they are visible at night, when their light does not pass by the Sun. This experiment was performed by Arthur Eddington during the total solar eclipse of 1919, and the result was found to be consistent, within experimental error, with the prediction of general relativity. This experiment caught the public fancy and was

The bending of starlight by the Sun

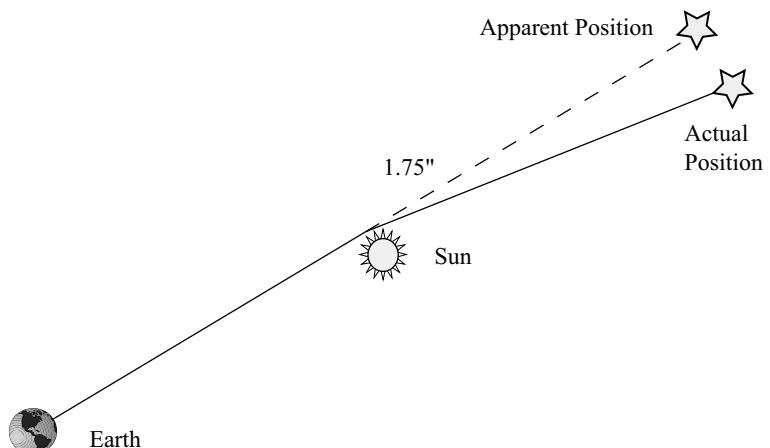


Fig. 8.10 Bending of starlight as it passes through the gravitational field of the Sun. The angle is greatly exaggerated in the figure; the actual angle is only 1.75 seconds of arc.

The orbit of Mercury explained

responsible, more than anything else, for the elevation of Einstein to the exalted status of a popular hero.

General relativity also predicts that the orbits of the planets will differ from the predictions of Newtonian gravity in flat space-time. It had long been known that there was a discrepancy in the orbit of Mercury of 43 seconds of arc per century, after its motion was computed and corrected for perturbations due to other planets. This is a very small residual, but it was well beyond the limits of the measurements of the late 19th century. The cause of this discrepancy could not be easily explained within the context of Newtonian theory. One of the first problems that Einstein tackled with his new equations was the effect of curved space-time on orbits. To his delight, Einstein found that the curvature of space-time near the Sun accounted for the mysterious deviations in Mercury's orbit. General relativity predicted that the difference in the motion of Mercury due to the curvature of space-time would be precisely 43 seconds of arc per century. This does not prove that general relativity is correct, but it furnishes a simple explanation that accounts exactly for an observed datum; that fact by itself provides a powerful motivation for accepting the theory. Since the late 20th century this type of measurement has been greatly refined with the help of space technology. Radar waves have been bounced off Venus and Mercury, determining their positions and orbits to great accuracy. Communications with spacecraft, particularly with the Viking Lander while it was on the surface of Mars, made it possible to measure the distance to that planet to within centimeters. These very exact measurements of planetary orbits make it possible to map the gravitational field of the Sun to extremely high precision. The results are all consistent with the predictions of general relativity.

The gravitational redshift and time dilation discussed earlier also provide a means to check the theory. The effect in the extremely weak field of the Earth is quite small, and clever experimentalists and very sensitive instruments are required, but the predictions of the equivalence principle have been verified. It is also possible, barely, to use the light

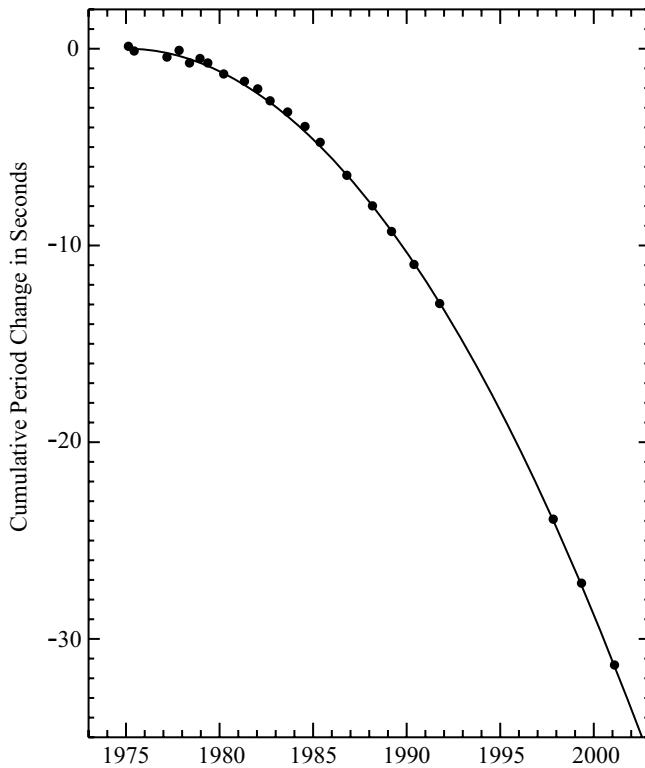


Fig. 8.11 The binary pulsar 1913+16 provides evidence for the existence of gravitational radiation. The loss of orbital energy to gravity waves causes the orbital period to change with time. The solid line is the prediction from general relativity; the dots are the observations. (Adapted from Weisberg and Taylor, 2003.)

from white dwarfs to test this aspect of the theory. The gravitational field of a typical white dwarf is sufficiently strong that the gravitational redshifting of photons departing its surface can be observed, with, of course, great difficulty; the shifts agree with the predictions of general relativity. The applicability of the theory for objects at astronomical distances is yet another confirmation that the physics we develop on Earth is valid for the universe as a whole. White dwarfs, many of which are near enough and bright enough to be observed easily, thus provide some quite tangible evidence for general relativity.

Recently, a very interesting test of general relativity has become possible with the discovery of binary pulsars. General relativity predicts that two compact and massive objects in orbit about one another will radiate gravitational waves. The loss of energy to those waves will cause the orbits to decay; the objects will gradually spiral inward. Since a pulsar acts like a very precise transmitting clock, we can follow its motion in its orbit very closely and can determine the orbital period to almost fantastically high precision. Over years of watching one such binary pulsar, astronomers have found that the orbit is changing in exactly the way predicted by general relativity. The binary pulsar thus serves as an indirect detection of gravitational radiation. It was for the discovery of the first binary pulsar and its very significant consequences that Joseph Taylor and Russell Hulse won the 1993 Nobel Prize in Physics.

Evidence for gravitational waves

Fig. 8.12 The LIGO gravity-wave detector at Hanford, Washington. LIGO, basically a huge Michelson–Morley interferometer, consists of two 4-km long evacuated pipes at right angles. Gravity waves can cause small differences in light-travel time down the pipes, which can be detected as shifts in the interference fringes when the beams recombine. (Caltech/NSF/LIGO.)



Of course, we would prefer to have a direct detection of gravitational waves, and it is possible, in principle, to detect such waves. Since space-time curvature is changing locally in a gravitational wave, a tidal force is induced in any physical object through which the wave passes. One promising technology for detecting these tidal forces is the laser interferometer, a device in which a beam splitter creates two light beams that are sent for a round trip along perpendicular paths. If the distance in either of the two directions changes in time due to a stretching or compression by a gravitational wave, the interference pattern changes when the light beams are recombined. This is, of course, the Michelson–Morley experiment adapted to the search for gravitational radiation. The experiment is currently just at the limits of technology, as the effect at the scale of any laboratory-bound system is incredibly tiny, on the order of one part in 10^{20} . With such a small effect, we would have a much better chance of detection if we could build an apparatus with an extremely long baseline.

Such a device has now been constructed. The Laser Interferometer Gravitational Wave Observatory (LIGO), which began operation in 2002, consists of two large laser interferometers, one located at Hanford, Washington and the other at Livingston, Louisiana. Each interferometer consists of perpendicular 4-foot diameter vacuum pipes, 4 kilometers in length, arranged in the shape of an “L.” Test masses with mirrored surfaces hang at the end of each arm. Laser beams travel back and forth down the pipe multiple times, creating a very long effective baseline. Because such a system is subject to many types of noise, for example, vibrations in the Earth, it is necessary to compare the signals at widely separated interferometers to look for common effects that might be due to gravity waves. LIGO is part of an international network of planned and operating gravitational wave observatories, including one in Italy, one in Germany, and one in Japan.

What might a gravitational wave observatory see? We have found indirect evidence for gravitational radiation from a binary pulsar. This radiation is not directly detectable by LIGO, but the universe must contain many binary pulsars. Eventually such a system’s mutual orbit

will shrink so much that the two neutron stars will spiral into one another. Such a cataclysmic collision would generate very strong gravitational radiation, producing effects which, though still small by ordinary standards, should be detectable on Earth; depending, of course, on the distance to the binary. The collision of two neutron stars is hardly an everyday occurrence in the neighborhood of the Sun, so we still need great sensitivity in order to be able to sample a very large volume of space, possibly to a radius of as much as one billion lightyears.

General relativity has shown us how matter and the geometry of space-time are related. It has provided us with new insights into the workings of the universe, and has predicted some remarkable new phenomena. General relativity has a profound impact on cosmology, fundamentally altering our view of the relationship of space and time. Yet how does that affect the humble stars and galaxies? Light may bend, gravitational waves ripple about, but it might appear that the mechanisms work pretty much as they always did in Newton's grand clockwork. However, we cannot make a radical change in the underlying paradigm of the universe without finding unexpected consequences for things once thought to be quite ordinary. Before returning to cosmological models, we will examine one of the most extreme consequences of general relativity, the black hole, whose properties are almost beyond our imaginations.

Chapter Summary

Special relativity showed that the absolute space and time of Newtonian physics could be only an approximation to their true nature. However, the special theory of relativity is incapable of explaining gravity because it assumes the existence of inertial frames; it does not explain how inertial frames are to be determined. Mach's principle, which states that the distribution of matter determines space and time, suggests that matter is related to the definition of inertial frames, but Mach never elucidated any means by which this might happen. General relativity attacks this problem and consequently discovers that gravity is related to geometry. The equivalence principle states that gravitational and inertial mass are equivalent; this is the fundamental basis for the general theory of relativity. The strict equivalence between gravity and inertial acceleration means that freefalling frames are completely equivalent to inertial frames. Geometry is related to matter and energy through Einstein's equation. In general relativity, the space-time geometry determines freefalling (inertial, geodesic) worldlines, and the geodesics specify how matter moves. Matter, in turn, tells space-time how to curve. The metric equation supplies a formalism for the space-time interval in general geometries, not just in

the Minkowskian (flat) space-time of special relativity. Matter and energy determine inertial frames, but within an inertial frame there is no influence by any outside matter. Thus Mach's principle is present more in spirit than in actuality in the general theory of relativity.

Tidal forces prevent a perfect equivalence of freefall and gravity. If the gravitational field diverges over the extent of an object, the various parts of the object will be pulled by forces of different strengths or in different directions. These differential effects are known as tidal forces. The equivalence principle requires only that the size of the freefalling frame be sufficiently small that tidal forces are negligible in order for the frame to be inertial.

General relativity predicts the bending of light by gravity, gravitational time dilation and length contraction, gravitational redshifts and blueshifts, the precession of Mercury's orbit, and the existence of gravitational radiation. All these effects have been measured, and gravitational radiation has been observed indirectly via the decay of the orbits of binary pulsars. The LIGO project's goals are more ambitious: to detect gravitational radiation directly using Michelson interferometers.

Key Term Definitions

freefall Unrestrained motion under the influence of a gravitational field.

Newtonian equivalence principle The principle that the laws of mechanics are the same in inertial and freefalling frames of reference. This implies that gravitational mass and inertial mass are equivalent.

Einstein equivalence principle The principle that *all* physical laws, not just those of mechanics, are the same in all inertial and freely falling frames of reference.

gravitational redshift A shift in the frequency of a photon to lower energy as it climbs out of a gravitational field.

Mach's principle The principle, elucidated by Ernst Mach, that the distribution of matter in the universe determines local inertial frames.

flat geometry Geometry in which the curvature is zero; ordinary Euclidean geometry.

spherical geometry A geometry that has positive constant curvature.

hyperbolic geometry A geometry that has negative constant curvature. Hyperbolic geometries cannot be fully visualized, because a two-dimensional hyperbolic geometry cannot be embedded in a three-dimensional Euclidean space. However, the central point of a saddle, that point at which curvature goes both “uphill” and “downhill,” provides a local representation.

metric equation The expression that describes how to compute the distance between two infinitesimally

separated points (or events) in a given geometry. Also called simply the “metric.”

metric coefficient The functions in the metric that multiply the coordinate differentials (for example, Δx) to convert these differentials into physical distances.

geodesic In geometry, that path between two points or events which is an extremum in length. In some geometries, such as Euclidean, the geodesics are the shortest paths, whereas in others, such as in the space-time geometries appropriate to general relativity, the geodesics are the longest paths.

tidal force In Newtonian gravity, the net force on an extended body due to a difference in gravitational force from one region of the body to another. In general relativity, a force arising when nearby geodesics diverge in space-time, because the world-lines of all parts of an extended body cannot travel along a single geodesic.

Riemannian geometry A generalized geometry that has the property of being locally flat; that is, in a sufficiently small region, a Riemannian geometry can be approximated by a Euclidean or Minkowskian geometry.

gravitational radiation The emission of gravitational waves by a gravitational field that changes in time. Also, the gravitational waves so radiated.

gravitational wave A propagating ripple of space-time curvature that travels at the speed of light.

Review Questions

- (8.1) What is “special” about special relativity, and what is “general” about general relativity?
- (8.2) Make a table of the properties of the three homogeneous and isotropic geometries we have studied: spherical, flat, and hyperbolic. Include answers to: (1) is it finite or infinite? (2) how does the circumference of circles relate to the radius of the circle? (3) what is the sum of the angles inside a triangle? (4) given a line, how many parallel lines are there through another point not on that line (the parallel-line postulate)?
- (8.3) State in your own words the meaning of Mach’s principle by considering the following thought experiment: the universe contains two observers who are initially at rest with respect to each other. Newton says that if the first observer accelerates away from the second, the first observer will feel a force, while the second will not. Why would Newton say this? What would happen according to Mach? What would Newton and Mach say if these two observers were the only objects in the universe?
- (8.4) What is the difference between the Newtonian version of the equivalence principle (the weak form) and the Einstein equivalence principle (the strong form)?
- (8.5) Consider the following experiment (which you can actually perform): Obtain a spring scale such as a typical bathroom scale, place it in an elevator, and stand on it. Note the *exact* value when the elevator is at rest. Now ride up several floors. As the elevator starts up, there is an acceleration upward. Note how the reading on the spring scale changes. Next ride down. When the elevator starts down, note how the reading changes. Once the elevator reaches a constant velocity up or down, note the reading of the scale. What do you predict these various readings would be?
- (8.6) A space station in deep space is spun like a giant wheel to produce centrifugal force so the occupants experience artificial gravity of one g . How does a clock at the rim of the space station compare with one residing at the hub? What does this say about the behavior of a clock sitting on the surface of a planet with a surface gravity of one g ?
- (8.7) Define *geodesic*. What does this mathematical quantity have to do with frames of reference?
- (8.8) (More challenging.) (a) Imagine a point source with the mass of Jupiter, 1.9×10^{27} kg, at a distance of 4.2×10^8 meters from a spherical object that has a diameter of 3,640 meters and a mass of 8.9×10^{22} kg. Consider a location on the surface of the sphere that is closest to the point mass. Now consider the location on the sphere that is exactly opposite the first location. (Refer to Figure 8.9(a) for a diagram of the situation.) What is the *difference* in the gravitational force between these two locations? How does this tidal force compare, as a percentage, with the gravitational force between the sphere (more precisely, the center of the sphere) and the point mass? You may use strictly Newtonian gravity for your answers.
- (b) The figures in this problem correspond approximately to Jupiter’s moon Io. Io is the only body in the solar system other than the Earth that is known to have active volcanoes. (Many other objects have extinct or at best dormant volcanoes, but eruptions have been photographed on Io by spacecraft.) The ultimate source of the energy of the Earth’s volcanoes is radioactive decay of uranium and thorium deep in the planet’s core. Io probably lacks such a source. Does this problem suggest to you a possible energy source for the vulcanism of Io?
- (8.9) Describe two distinct experimental tests of general relativity. Explain how the results distinguish between Newtonian gravity and general relativity.
- (8.10) Explain in your own words what a gravitational wave is. At what speed do such waves propagate?
- (8.11) A spaceship is coasting in orbit around a planet. A second spaceship sits motionless on the launch pad. The two ships define frames that are accelerated with respect to each other, yet both might be regarded as inertial frames. Explain.
- (8.12) Under what circumstances is Newtonian mechanics valid? Does the development of general relativity mean that Newtonian theory is useless, or is an unacceptable scientific theory? Why or why not?

This page intentionally left blank

Black Holes

9

Key Terms:

There's always a hole in theories somewhere, if you look close enough.

—Mark Twain, *Tom Sawyer Abroad*

Schwarzschild's solution

The death of supermassive stars must result in collapse; no known force can resist gravity in such stars once their nuclear fires have died. The result of this inevitable collapse is known as a **black hole**. The term black hole is nowadays bandied about so much, in science-fiction novels and movies, as well as in the occasional general-science articles of newspapers and magazines, that it would be difficult for any reasonably literate person to be unaware of the expression. Yet few understand why black holes exist, or what their properties really are. They are often imagined as some sort of monstrous and voracious maw, devouring anything that comes too near, even light and energy. While accurate in some respects, such a notion falls far short of a complete description of black holes and their remarkable properties.

The black hole is an extreme consequence of Einstein's theory of general relativity, but the possibility of something like it exists even within the Newtonian theory of gravity. Escape from any star or planet requires a velocity large enough to overcome the gravitational pull at the object's surface. This velocity is known as the **escape velocity**. In Newtonian gravity, the escape velocity from a spherical object of radius R and mass M is

$$v_{\text{esc}} = \sqrt{2GM/R}. \quad (9.1)$$

For the Earth, the escape velocity is about 11 km s^{-1} . What if there existed a star with an escape velocity equal to the speed of light? No light could leave its surface; it would be a dark star. Light shining from the surface of this star might climb up, but like a ball thrown into the air, it must eventually reverse and fall back down. When this idea was first proposed, it was not known that the speed of light *in vacuo* is the ultimate speed limit, but with that additional fact, it is easy to conclude that *nothing* could escape from such a star. Setting $v_{\text{esc}} = c$ in equation (9.1) and solving for the radius gives $R = 2GM/c^2$. For a star with the mass of our Sun, this radius is about 3 kilometers; the Newtonian dark star is very compact indeed.

- black hole
- escape velocity
- Schwarzschild radius
- event horizon
- singularity
- coordinate singularity
- cosmic censorship
- quantum gravity
- Birkhoff's theorem
- photon sphere
- gravitational lens
- no-hair theorem
- Kerr metric
- static surface
- ergosphere
- Hawking radiation
- uncertainty principle
- virtual particle
- black hole thermodynamics
- accretion disk
- quasar (QSO)
- active galaxy
- active galactic nucleus (AGN)

The Newtonian dark star

The story of the general-relativistic black hole begins late in 1916. Despite the great complexity of the Einstein equations, Karl Schwarzschild found one of the first solutions almost immediately after Einstein published his final results. Schwarzschild assumed a perfectly spherical, stationary ball of mass M , surrounded by a vacuum. This is not a bad approximation to a star; the Sun rotates slowly and is very close to spherical, and as far as we know, the Sun is a typical star. Moreover, the space immediately beyond the Sun and most stars is a decent approximation to a vacuum. Schwarzschild then solved Einstein's equations to compute the space-time curvature in the *exterior* of the star. Such a solution consists of a specification of the geometry of space-time; as we have discussed, this description can be encapsulated in the metric coefficients, as indicated by equation (8.4).

Schwarzschild's assumptions greatly simplified the mathematics required. First, he was solving for the gravity in a vacuum outside the mass. This meant that he could set the stress-energy term $T^{\mu\nu}$ in Einstein's equation equal to zero and work only with the geometry term. Since he was considering the space around a spherical mass, Schwarzschild employed spherical spatial coordinates, consisting of a distance R from the center of the mass, as well as the inclination from the origin, expressed in terms of two angles such as altitude θ and azimuth ϕ . (The precise definition of the radial distance is slightly more complicated than this, but the details need not concern us here.) The gravity arising from such a star must be spherically symmetric; that is, it should depend only on the distance from the star. Thus it was possible to ignore the angular terms, another considerable simplification for Schwarzschild. Finally, the star and its gravitational field are unchanging in time. This implies that the metric terms cannot depend on time, assuming that the time coordinate is sensibly chosen. The time coordinate Schwarzschild employed was a very reasonable choice; it corresponds to the time measured by an observer very far from the central mass, where gravity's effects diminish toward zero. With all these simplifications, Schwarzschild obtained his metric,

$$\Delta s^2 = \left(1 - \frac{2GM}{c^2R}\right) c^2 \Delta t^2 - \frac{\Delta R^2}{\left(1 - \frac{2GM}{c^2R}\right)} - R^2(\Delta\theta^2 + \sin^2\theta\Delta\phi^2). \quad (9.2)$$

This is a full general-relativistic metric, or space-time interval, in all its glory. The Schwarzschild metric is similar to the familiar flat space-time interval of special relativity, as written in spherical coordinates, but it is modified by the appearance of the metric coefficients, which vary only with R . These new functions affect only the time and the radial measurements; the angular terms are unchanged from ordinary flat space, and we can ignore them henceforth. We can interpret these metric coordinates in terms familiar from our previous study of general relativity; the Schwarzschild coefficients of Δt^2 and ΔR^2 respectively account for gravitational time dilation and length contraction. Keep in mind that this solution is valid in a vacuum outside *any* spherical body

The Schwarzschild metric

of mass M and radius R . It does not, however, hold in the interior of the body.

The combination

$$2GM/c^2 \equiv R_s \quad (9.3)$$

appears in both of the new metric coefficients. This expression is very important, and will turn out to be intimately linked with many of the unusual properties of black holes. It has units of length and is called the **Schwarzschild radius**, R_s . Because c^2 is large, R_s will be extremely small unless M is also large. For the mass of the Earth, R_s is equal to about a centimeter. Does this imply that the matter within a centimeter from the center of the Earth is within the Earth's Schwarzschild radius? No, because the Schwarzschild solution only applies *outside* a mass. The Earth is filled with mass whose distribution is a function of radius; where mass is present, it is necessary to solve the Einstein equations with the stress-energy term present.

Because the radius of the Earth is so much larger than the Schwarzschild radius, the metric expression $2GM/c^2R = R_s/R$ will be very tiny for the gravitational field surrounding the Earth. This means that the modifications to ordinary flat space and space-time will be equally small. Consequently, space and space-time in the vicinity of the Earth are curved very little, although this small curvature still accounts for the gravitational field we experience. The major effect on space-time around the Earth, the Sun, or any other spherical object that is large compared to its Schwarzschild radius, occurs through the metric coefficient of the time coordinate, due to the presence of the speed of light in the expression $(1 - R_s/R)c^2\Delta t^2$. The contribution of the radial coefficient is much smaller, with a correspondingly minuscule curvature to the *space* around the Earth; thus space in our vicinity remains very nearly the familiar Euclidean. In the relativistic view, we can say that the Earth's gravity is mainly due to time curvature. This is true for nearly all ordinary objects; even for neutron stars the correction is still modest. But what about an object whose radius R is comparable to $2GM/c^2$? The coefficient of Δt shrinks toward zero, while that of ΔR becomes enormous. The general-relativistic properties of such a compact star become increasingly evident.

As the simplest example of relativistic behavior near the Schwarzschild radius, consider the gravitational time dilation of a clock located at some radius R . We have learned the general rule that a clock in a stronger gravitational field runs slower than an identical clock at a location where the field is weaker; how can we make this more exact? It is the metric that enables us to compare the rate of this clock to one at a great distance. The time component t in the metric equation corresponds to that measured by a clock at infinity. The metric coefficient gives the time-dilation factor, namely,

$$\Delta\tau = \sqrt{1 - R_s/R} \quad \Delta t. \quad (9.4)$$

The ratio R_s/R is a measure of the strength of the gravitational field

Gravitational time dilation

For example, a clock just outside the Schwarzschild radius, at $R = 1.33R_s$, is ticking at half the rate of the clock at infinity. Only one hour passes here for every two of the distant observer.

The Schwarzschild metric affects not only time, but also space. What happens to a standard length, that is, a ruler, in the Schwarzschild metric? The gravitational length contraction is determined by the radial metric coefficient. The length of a stationary ruler at radius R is related to the length of a ruler at infinity by

$$L = L_\infty / \sqrt{1 - R_s/R}, \quad (9.5)$$

where L is the length of the ruler located at distance R , and L_∞ denotes the rest length measured by the distant observer. A meter stick located just outside the black hole at $R = 1.33R_s$ is only half a meter in length, as measured by the distant observer.

The metric affects not only space-time, but also anything traveling through space-time, including light. One of the most important consequences of the effect of the metric upon the propagation of light is the *gravitational redshift*, which, as we have learned, is a consequence of the equivalence principle. Now that we have a specific metric, we can compute an explicit formula for the corresponding gravitational redshift. Redshift is *defined* to be

$$z = \frac{\lambda_{\text{rec}} - \lambda_{\text{em}}}{\lambda_{\text{em}}} = \frac{\lambda_{\text{rec}}}{\lambda_{\text{em}}} - 1, \quad (9.6)$$

where λ_{rec} is the wavelength of the light received at the detector and λ_{em} is the standard wavelength, that is, the wavelength measured in the rest frame of the emitter. Wavelength is a length, and will be contracted by the Schwarzschild gravitational field in the same way as any other length. Thus the gravitational redshift of a photon emitted at a distance R from the center of a compact object and received at infinity is simply

$$z = \frac{1}{\sqrt{1 - R_s/R}} - 1. \quad (9.7)$$

Since the Schwarzschild solution is valid only outside of a star, this formula holds when R is greater than the star's radius. Although we derived this result from consideration of length contraction, the identical result could be obtained from time dilation, because longer-wavelength radiation has a lower frequency, according to the relation $\nu = c/\lambda$, and frequency is simply an inverse time interval.

Gravitational length contraction and time dilation occur in any gravitational field. But what if an object's radius were equal to the Schwarzschild radius? At $R = R_s$ the coefficient of Δt^2 becomes zero, and that of ΔR^2 becomes infinite. Does this mean that space-time has broken down? For long after Schwarzschild presented his solution, it was believed that the solution simply was not applicable for so small a radius, and therefore no physical object could ever be smaller than its Schwarzschild radius. It took quite a while for scientists to realize that the solution does not fail. Instead, what fails at the Schwarzschild radius

Gravitational length contraction

Gravitational redshift

is our choice of coordinates R and t . It is an artificial failure, similar to what would happen if we decided to measure temperature in terms of the inverse of degrees Celsius. On such a temperature scale, the freezing point of water still exists, but we have chosen a particularly inappropriate way to measure it, a marker that becomes infinite at this particular point. Once it was realized that the Schwarzschild solution does *not* break down at the Schwarzschild radius, it became necessary to consider the consequences of a star that had collapsed to such an extent. These are the objects now known as black holes.

Properties of black holes

It seems rather coincidental that the Schwarzschild radius is the same as the radius derived previously for the Newtonian dark star. Perhaps, since Newtonian gravitation is valid to a good approximation, we should have expected something not too far from its prediction. But the black hole is a much more interesting and exotic object than is the dark star, and insistence upon thinking about the black hole as if it were a Newtonian dark star will lead to misunderstanding of the essential properties of the black hole.

Why is every spherical object not a black hole? Because the Schwarzschild radius lies well within the outer surface of any normal object, even a neutron star. For example, the Schwarzschild radius of the Sun is approximately three kilometers, compared to a solar radius of almost a million kilometers. The Schwarzschild radius of the Earth is less than a centimeter. As emphasized above, the Schwarzschild solution applies *only* in the empty space to the exterior of the sphere; if the Schwarzschild radius is less than the physical radius of the body, then Schwarzschild's solution is irrelevant within the body. The metric inside a star is not a Schwarzschild metric, but a different metric that takes into account the presence of the matter which generates the gravitational field. Only if the object has collapsed completely and disappeared beneath its Schwarzschild radius can a black hole be formed.

At the Schwarzschild radius, the coefficient of the time interval Δt in the Schwarzschild metric goes to zero. Therefore, the time interval itself, which is the proper time divided by this coefficient, becomes infinite; clocks stop. Similarly, radial intervals fall to zero, the ultimate length contraction. These effects are a consequence of our choice of coordinates, and coordinates themselves are not absolute even in Newtonian physics. Nevertheless, the time dilation, length contraction, and other relativistic effects that depend directly upon the metric coefficients, are real physical phenomena that can be measured with sensitive instruments. As is true for any massive object, the gravitational field near the black hole is stronger at small radius than it is far away, and so light climbing from close to the object suffers a gravitational redshift. In the case of a black hole, any light sent from the Schwarzschild radius is *infinitely* redshifted. The sphere defined by the Schwarzschild radius thus represents a surface

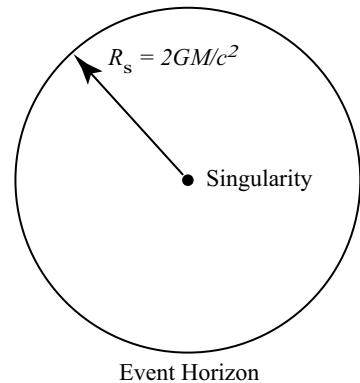


Fig. 9.1 Components of a stationary black hole. The event horizon, located at the Schwarzschild radius, defines the size of the black hole. The singularity at the center is the point at which all ingoing worldlines end, and matter is crushed to infinite density.

The event horizon

from which light cannot travel to an external observer. An observer outside this surface can never see within it; the interior of the black hole is forever cut off from communicating with the rest of the universe. Events inside the black hole can have no causal contact with events to the exterior. This causal boundary between the inside and the outside of a black hole is an **event horizon**. It is the surface from which no light or other signal can ever escape. Thus the Schwarzschild radius marks the event horizon of the black hole.

From outside a black hole, the event horizon seems to be a special location. What would happen if an advanced civilization were to launch a probe toward a black hole? To the observers watching from a safe, far distance, the infalling probe's clock slows down; radio signals from the probe come at increasingly longer wavelengths due to the gravitational redshift. The probe approaches closer and closer to the horizon, but the distant observers never see it cross over into the hole. Time seems to come to a halt for the probe, and the redshift of its radio beacon goes to infinity, as measured by the faraway astronomers. At some point the last, highly redshifted signal from the probe is heard, and then nothing more. The probe disappears forever.

Does this mean that the probe is destroyed upon reaching the horizon? No; these strange effects, such as the freezing of time for the probe, are artifacts of the space and time coordinates of the *external* observers. The Schwarzschild radius is not a true **singularity** in the metric, a place where tangible, physical quantities such as pressure or density reach infinity, but rather it is a **coordinate singularity**, a point at which our choice of coordinate system fails. However, only the coordinate system defined by the observers at infinity fails; a coordinate system falling freely with the probe remains valid, and indicates no changes in time or length values. Time and space seem normal to the probe, even at the horizon.

Extreme time dilations and length contractions are not unique to general relativity; an example from special relativity would be a spaceship accelerating toward the speed of light. To an observer at rest, the relativistic spaceship's clocks would seem to slow toward a halt, while meter sticks aboard the spaceship would shrink toward zero length. Yet the occupants of the spaceship would see nothing strange occurring. Similarly, to the ill-fated probe approaching the horizon of the black hole, nothing unusual occurs; physics continues to appear perfectly normal. This follows from the nature of space-time and the equivalence principle. Even near a black hole, a sufficiently restricted, freely falling frame must be equivalent to any other inertial frame.

However, there are other effects near the Schwarzschild radius that might affect an infalling probe. Since the probe is an extended body, it is subject to tidal forces; hence it might well be in danger near the event horizon. Tidal forces result when the gravitational force varies over a body. The gravitational field near a black hole increases so rapidly as the event horizon is approached that the part of the probe closest to the black hole might experience a substantially larger gravitational

The fate of a probe falling into a black hole

Physical vs. coordinate singularity

Tidal forces around a black hole

force than would more distant parts. The curvature of space near the black hole is comparable to the scale of its horizon; because inertial frames must be locally flat, near a black hole these local inertial frames must be very restricted indeed. Tidal forces near a black hole will be large for bodies whose size is not extremely small in comparison to the radius of the horizon. If a spaceship were to fall into a stellar-sized black hole, with a Schwarzschild radius of approximately 3 km, it and its occupants would be torn apart by tidal forces. For such a small black hole, any spaceship that could accommodate human-sized bodies would occupy a spatial volume that cannot be approximated by flat space-time in the vicinity of the event horizon. But a black hole need not always chew its food before swallowing it; if a spaceship fell into a black hole a million times more massive than the Sun, with a horizon radius correspondingly a million times greater than the Schwarzschild radius of the Sun, the volume surrounding the spaceship would be fairly well approximated locally by a flat space-time, and it would not experience strong tidal forces. In such a case, the crew would scarcely notice their passage across the event horizon. An even larger black hole, such as might occur from the collapse of the core of a galaxy, would produce reasonably small tidal forces even at the scale of an object as large as a star.

Thus, if a probe fell into a sufficiently large black hole, it would feel no ill effects as it crossed the horizon. What might be found within the horizon of the black hole? Although the infalling probe could never return data from the inside, the Einstein equations still hold and still describe space-time. If we continue to use the time and space coordinates appropriate for an observer at a great distance from the hole, as we have in equation (9.2), we find that within the Schwarzschild radius, the metric function behaves in a peculiar manner. When $R < R_s$, the metric coefficient of the time coordinate becomes negative, while that for the radius becomes positive, in a reversal of their usual signs outside the horizon. This suggests that within the black hole, the time and space coordinates, as defined by the external observer, exchange their roles. Recall that material particles must have worldlines for which $\Delta s^2 > 0$. Outside the horizon, a worldline can be fixed in space, with $\Delta R = 0$, as it advances forward in time. On the other hand, no worldline could be fixed in time, while moving through space. Within the black hole, in contrast, if the particle's worldline remained at a fixed radius from the center, that is, $\Delta R = 0$, then the space-time interval would become negative, or spacelike, which is not allowed for a particle worldline. Therefore, within the Schwarzschild radius, it is impossible for a particle to orbit at a fixed radius; its radius *must* constantly decrease. The future, as it were, lies inward.

Let us consider this in terms of lightcones. At any event (R, t) in space-time, we can construct a lightcone, as illustrated in Figure 9.2. For example, we could position a particle at some location and let it emit a pulse of light at some instant in time. Far from the black hole, the lightcones are just as they would be in Minkowskian space-time.

Inside the horizon

Lightcones around a black hole

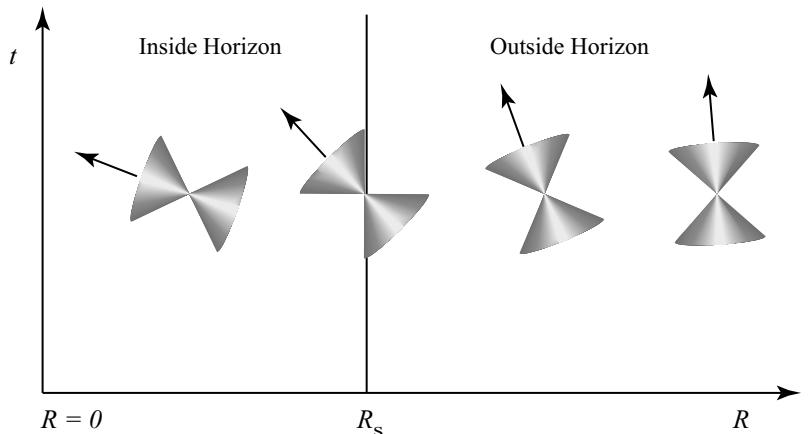


Fig. 9.2 Tilting of lightcones as a particle approaches a black hole. As the horizon nears, more and more of the future lies toward the hole. The horizon itself coincides with a lightlike (null) surface, and the future lies entirely inward. Inside the horizon, the external coordinates t and R have swapped roles, and the future lies in the direction of decreasing radius.

Nearer to the black hole, however, geodesic paths, including the lightlike paths followed by light rays, point more and more toward the hole, that is, toward $R = 0$. This means that near the event horizon, much of the light emanating from the emitting particle would fall into the black hole. As the horizon is approached, an ever-increasing fraction of the particle's possible future worldlines, which must be contained within its future lightcone, would point toward the hole. In other words, the lightcones begin to tilt toward the hole. At the horizon, *all* of the particle's future will lie inward; one edge of the lightcone will coincide with the horizon. This edge would describe a light beam directed straight outward, but frozen forever exactly at the horizon. Once inside the horizon, the lightcone is even further tipped over. The future is directed toward smaller R , the past toward larger R . This is another way to look at the interchange of time and space coordinates; out here we may say that the future lies with greater values of time t . In there the future lies toward smaller values of the radius R . A worldline in the interior of the hole could move in the $+t$ or $-t$ direction, but that still does not permit time travel, because a worldline can never emerge outside of the horizon, where t is once again the usual time coordinate.

A useful way to visualize this phenomenon is to imagine that space-time is like water; a black hole is analogous to a drain. Objects falling radially toward the black hole are like boats floating unpowered in the water, moving with the current. Far from the black hole, our boat drifts very slowly toward the horizon; but the water, and hence the boat, gains speed the closer we approach the hole. If we wish to avoid falling down the drain, we must turn on our motor and aim away from the hole. There comes a point, however, at which the water is flowing inward faster than the motor can drive the boat. In this analogy, we can think of space-time itself drawn inward at an increasing rate by the gravitational pull of the black hole. At the horizon, space-time flows inward at the speed of light, so that a light ray emitted against the flow can, at best, stand still. Inside the horizon, space-time flows at a speed faster than that of

light, so even light cannot move outward anymore. (To forestall protests that nothing can move faster than light, we emphasize that this is an analogy; moreover, the motion of space-time, as we are describing it, carries no information and hence its superluminal speed cannot violate causality.)

Once inside the horizon, the radius of any particle's orbit must inexorably decrease, and any particle that crosses the horizon must eventually fall into the center. At the center our metric equation once again fails, here because $R = 0$. This time, however, the failure is real, and unavoidable by a change of the coordinate system. At the center of the hole lies the true singularity, the point at which density becomes infinite. Any particle that crosses the event horizon is doomed, since it must fall toward the infinite crush at the center. The exact proper time required for infall to the singularity depends upon several factors, such as the path taken, but it is approximately equal to the time for light to travel a distance equal to the Schwarzschild radius. The larger the black hole, the longer this time is. If a particle fell straight into a solar-mass black hole, it would reach the singularity in roughly ten microseconds of proper time. Similarly, infall from the horizon of a black hole of mass $10^8 M_\odot$, such as might inhabit the cores of many galaxies, would take only 16 minutes.

Whatever may be going on at the singularity of a black hole does not matter to the external universe. The singularity is surrounded by the event horizon, and hence no information or signals from the singularity can ever emerge. If an astronaut were to venture into a black hole, giving up his life in order to see the singularity, the sacrifice would be in vain. We can prove mathematically that it is impossible to observe light rays traveling from the singularity in the Schwarzschild solution; even within the event horizon, light rays cannot move toward *any* larger R , including the astronaut's position as he falls, so the singularity is invisible even from inside the horizon. But what about singularities that might exist in other solutions, including those that we have not yet discovered? Might some solutions contain *naked singularities*, bare singularities that have no event horizons to shroud them? The conjecture that no singularities can ever be seen because they must be surrounded by event horizons is known as the **cosmic censorship** hypothesis: there are no naked singularities. This proposition holds that whenever a singularity forms, it will do so within the confines of a shielding horizon; thus whatever the properties of a singularity may be, they can have no effect on the rest of the universe. Though no realistic exceptions are known, the cosmic-censorship hypothesis has not yet been proven; it is based on experience to date with the Einstein equations, and on our expectations about how the universe should work.

Is the density at the singularity truly infinite? Many scientists do not believe that infinite density can exist in the physical universe. We know, for example, that the general theory of relativity has never been made fully consistent with quantum mechanics, the other triumph of the physics of the first half of the 20th century. The required theory would

Physical singularities are hidden

provide an explanation of **quantum gravity**, but no such theory has yet been developed. It is likely that there is a point at which Einstein's equations break down as a suitable description of the universe, and it may be that quantum effects prevent a literal singularity. In any case, it is probable that our current notions of particles, and perhaps even our conceptions of space and time themselves, fail at a singularity. Even if such an effect occurs, however, the center of a black hole represents the highest density possible in the universe.

Perhaps it is because of the singularity that black holes are sometimes said to be the “densest things in the universe.” Black holes may be very dense, but they need not be so. The black hole is not the singularity *per se*, but the volume of space surrounded by the event horizon. Just as for any other spherical object, a black hole’s density is proportional to its mass, divided by the Schwarzschild radius cubed. The radius of the black hole is itself proportional to the mass of the hole; hence the volume of the hole is proportional to the cube of its mass. Thus the *average* density of a black hole goes as the inverse of the mass squared; the more massive the black hole, the less dense it is. Specifically,

$$\rho_{\text{bh}} = \frac{3M}{4\pi R_s^3} = \frac{3c^6}{32\pi G^3 M^2} \propto \frac{1}{M^2}. \quad (9.8)$$

The density ρ_{bh} indicates how much a mass M would have to be compressed to create a black hole. For example, the Sun would have to be compacted to a radius of 3 km to form a black hole. Since the Sun has a mass of 2×10^{30} kg, its density as a black hole would be about 10^{19} kg per cubic meter. This is indeed fabulously dense, far beyond the imagination of any of us, and considerably greater than the density even of an atomic nucleus, which is typically about 10^{17} kg per cubic meter. On the other hand, a black hole 100 million times as massive as the Sun, with a radius proportionally larger, would have an average density approximately the same as that of water, hardly an unusually dense substance. If a black hole were created from the Milky Way galaxy by collapsing all its stars together, the entire galaxy would be contained within its horizon with the stars still well separated. In the most extreme limit, if the entire visible universe were in a black hole, its average density would be close to what is actually observed, about 10 hydrogen atoms per cubic meter. Thus, we could be located inside such a high-mass, low-density black hole without our immediate surroundings appearing in any way exotic. However, if we were in the interior of any black hole, we would have a limited (proper) time left to live, since nothing can stop the inevitable collapse into the central singularity.

Now let us return to the exterior of the black hole and ponder a few more of its properties. One of the most common misconceptions about black holes is that they possess some sort of supergravity power to draw distant objects into them. In reality, beyond its immediate vicinity a black hole has no more and no less gravitational pull than any other object of equal mass. At large distances from a black hole, its presence is not felt in any unusual manner; its gravitational field

The density of a black hole

is not qualitatively different from the gravitational field of any other object in the universe. All massive objects produce curvature in space-time. The unusual aspect of the field of the black hole is the strength of the curvature very near the event horizon. Far from the horizon, the gravitational field of the black hole is indistinguishable from the field of any other object of the same mass M . Just as in Newtonian gravity, in general relativity the gravitational field outside a spherically symmetric body behaves as if the whole mass were concentrated at the center. Moreover, **Birkhoff's theorem** states that the gravitational field outside any spherical object, black hole or ordinary star, cannot be affected by purely radial changes in the object. If the Sun were to collapse suddenly to a black hole, uniformly toward its center, we would certainly notice the absence of light, but its gravitational field at the distance of its planets would not change; the Earth would continue to orbit exactly as it does now. Indeed, the gravitational field would be unchanged right down to the former radius of the Sun. The bizarre effects of black-hole gravity would manifest themselves only in the new vacuum region between the original radius and R_s .

One such effect alters the properties of orbits around a black hole. In classical Newtonian gravity, it is always possible to orbit a gravitating body indefinitely, and arbitrarily closely to the body's surface, provided that no energy is lost to dissipation in an atmosphere. It is merely necessary to travel at a high enough speed, in a direction perpendicular to the radial direction, in order to balance the centrifugal and gravitational forces. In relativity, on the other hand, there is an ultimate speed limit, c ; nothing can orbit at a speed greater than that of light. Close to a black hole, there is a minimum radius within which gravity becomes so intense that no material object can orbit fast enough to resist infall. At distances smaller than this radius, there are no stable circular particle orbits at all. For a Schwarzschild black hole, this point occurs at three times the horizon radius,

$$R_{\min} = 3R_s. \quad (9.9)$$

Inside this point, a massive particle may fall in or fly out, but it cannot remain in orbit.

Space-time curvature also affects the path of light beams; near a black hole, the bending of light becomes extreme. At a distance of 1.5 times the Schwarzschild radius, the path of a light beam passing the hole on a trajectory perpendicular to the radial direction is so strongly bent that the beam turns and traces out a circular orbit around the black hole. This is called the *photon orbit*, and it occurs at

$$R_\gamma = 1.5R_s. \quad (9.10)$$

The radius of the photon orbit lies within the radius of the last stable particle orbit. The sphere defined by this radius R_γ is called the **photon sphere**. Within the photon sphere, not even light can remain in orbit, but must move radially inward or outward. A spaceship could not orbit at the photon sphere; it could, however, hover at this distance from the

Purely radial changes in a spherical star do not alter its external gravitational field

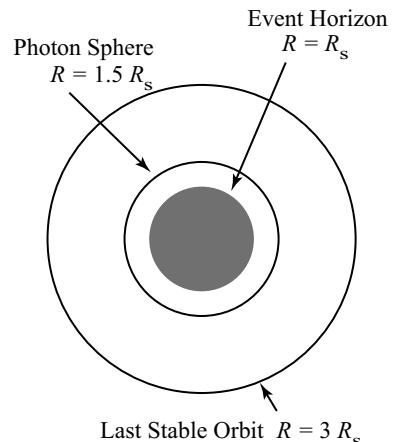


Fig. 9.3 Orbits around a Schwarzschild black hole. The photon sphere is the distance from the event horizon at which a light beam's path is bent into a circular orbit. The last stable orbit is the closest possible circular orbit for material particles.

The bending of light

black hole, albeit only by firing its rockets toward the hole. The crew aboard such a spaceship could look along the photon orbit and see the back of the ship, by the light curving around the hole!

The intense gravitational field near the black hole can produce many other interesting effects due to the bending of light. Any gravitational field causes light to deviate from the straight trajectories it would follow in an empty space-time. The effect is significant even for the field of the Sun, but is far more pronounced in the strong curvature of space-time around a black hole. If a black hole lay along the line of sight from the Earth to some distant galaxy, the light rays from that galaxy would be strongly bent and deflected as they passed near the black hole. Any sufficiently strong gravitational field, which might also result from a large mass concentration such as a very massive galaxy or a galaxy cluster, would have a similar effect, but the black hole is most effective at this phenomenon by virtue of the very strong gravity near its event horizon. Such bending and focusing of light by a gravitational field is called *gravitational lensing*, and the object that creates the image is called a **gravitational lens**. The gravitational bending of light is different from the bending in ordinary glass or plastic lenses. In these, light rays are bent by refraction, the change in the speed of the waves when they pass from one medium (air) to another (optical glass or plastic). Nevertheless, the bending of light rays by a gravitational field can, under the right conditions, also cause an image to form. Many examples of gravitational lensing have been observed, although none is definitely associated with a black hole.

Rotating black holes

Now, here, you see, it takes all the running you can do, to keep in the same place.

Lewis Carroll, *Through the Looking Glass*

So far our discussion of black holes has remained in terms of only one property, their mass. What other properties can black holes possess? Nearly every object in the universe rotates, so we would expect that any precursors to black holes would likely rotate. What happens when a rotating object collapses to a black hole? What if the star had a magnetic field, or an electric charge? What if the star is oddly shaped, or, as an even more exotic possibility, what if it were made of antimatter rather than matter? How would any of these things affect the black hole that is formed by the collapse of such a star? As remarkable as it may seem, the resulting black hole is very simple no matter how complex the object that forms it. The powerful singularity theorems of Roger Penrose and Stephen Hawking show that asymmetries or irregularities in the collapse will not prevent the formation of a singularity. It can also be proven that the only aspects of the precursor that are remembered by

a black hole are its *mass*, *electric charge*, and *angular momentum*. The space-time around a black hole will always settle down to a smooth, spherically symmetric configuration; any properties other than these three will produce nonspherical components of the field, which will be radiated away as gravitational waves. This theorem about the final state of the black hole is known among relativists as the **no-hair theorem**, from the saying “black holes have no hair”; it does not mean that they are giant eight-balls in space, but rather, that they have no detailed structure, or “hair,” emerging from the horizon that would violate its perfect spherical symmetry. For example, if more matter falls into one side of a black hole, the mass of the hole changes but the gravitational field adjusts to maintain the horizon’s spherical shape.

The no-hair theorem tells us that no matter how complex the properties of the precursor, only mass, electric charge, and angular momentum are preserved in the collapse to a black hole. Electric field lines emerge from a uniformly charged object in purely radial directions; hence the electric field is spherically symmetric and is remembered by the space-time geometry outside the black hole. However, it is unlikely that any black hole would actually maintain a net electric charge for very long. If a hole had any net charge, it would rapidly attract opposite charges until it was neutralized. Magnetic fields, on the other hand, are not spherically symmetric, and therefore any stellar magnetic field will be radiated away as electromagnetic waves. In this respect, a black hole is quite different from a neutron star. Neutron stars probably possess enormous magnetic fields, which account for a significant portion of the emissions from pulsars. Black holes have no magnetic fields of their own. Finally, if a black hole forms from a rotating object, which should be true in the majority of cases, it will remember the precursor’s original angular momentum. The Schwarzschild metric cannot describe a rotating black hole; for that we need a more general solution, the **Kerr metric**, published by Roy Kerr in 1963. The Kerr metric is an *exact* solution to Einstein’s equations for a rotating sphere, and it reduces to the Schwarzschild metric when the rotation is zero. When the rotation is not zero, however, the surrounding space-time is endowed with several new properties.

Very early in the history of general relativity, J. Lense and H. Thirring discovered what is now known as the *Lense–Thirring effect*, the *dragging of inertial frames* by rotating bodies. This phenomenon occurs for any rotating body, but it is extremely tiny for anything but a Kerr black hole. Frame dragging means that freefalling geodesics directed initially toward the center of the black hole will not fall straight along a purely radial path, but will spiral in the direction of the spin of the hole. In other words, the inertial frames near such a body partake of its rotation. If an explorer fell straight down toward the equator of a Kerr black hole from a great distance, it would feel to the *falling* observer that his path was straight and he was not rotating, but a far-off observer would see him spiraling inward as he neared the horizon. To him, on the other hand, it would seem that the distant stars would begin to rotate. Like a leaf

A black hole has no “hair”

Frame dragging

sucked into a vortex at the bottom of a waterfall, the explorer would be dragged into a spiral path by the flow of space-time. This is reminiscent of Mach's principle; the definition of what constitutes an inertial frame is influenced by the rotation of a nearby, dominant mass. In its own vicinity, a rotating body vies with the overall matter distribution of the universe to establish what constitutes a local inertial frame.

The rotation of a black hole also alters the event horizon. What was a single event horizon for a Schwarzschild black hole now splits into two surfaces. The inner surface, which is spherical and lies inside the usual Schwarzschild radius, is an event horizon, and it is similar to its Schwarzschild counterpart in that it represents the point of no return for an infalling particle. The radius of the event horizon of a Kerr black hole is given by

$$R_K = \frac{G}{c^2} \left(M + \sqrt{M^2 - a^2} \right), \quad (9.11)$$

where a is a measure of the spin angular momentum of the hole. If the hole is not rotating, then the angular momentum is zero, and the Kerr radius is equal to the Schwarzschild radius. As the spin of the black hole increases, the radius of the horizon shrinks. An even more interesting consequence of this equation is that there is a limit to the rotation speed even of a black hole. A black hole can exist only for $a \leq M$; a hole for which $a = M$ is said to be *maximally rotating*. At this limit, the Kerr horizon has shrunk to half the Schwarzschild value.

The static surface of a rotating black hole

The outer surface is called the **static surface**. It is oblate and touches the event horizon at the rotational poles of the black hole; it coincides with the Schwarzschild radius at the equator of the hole. This surface is called the static surface because at or inside this point nothing can remain static, that is, motionless, with respect to the spatial coordinates. If a spaceship slowly descends straight toward the hole while attempting to remain aligned with a distant star, the crew will find that in addition to firing an engine toward the hole to combat the pull of gravity, they must also aim against the direction of the hole's rotation, in order to maintain their position relative to the distant star. The effect becomes stronger as they near the static surface. Finally, at the static surface it is necessary to move at the speed of light opposite the hole's direction of spin to avoid being dragged with the rotation of the hole; that is, it is necessary to move as fast as possible, just to stand still! The black hole compels freefalling trajectories to participate in its rotation; freely falling lightcones tip increasingly toward the direction of rotation of the hole as the static surface is approached. Inside the static surface, the rotation of space-time is so great that not even light can resist being dragged around the hole. Unlike the event horizon, however, the static surface is not a one-way membrane; it is possible to pass through it from the outside and return to tell the tale.

The region between the horizon and the static surface is called the **ergosphere**. In principle, it is possible to extract energy from the er-

gosphere, a property from which its name is derived.¹ An advanced civilization might accomplish this by sending spaceships into an appropriate orbit just inside the static surface. The spaceships would eject something—nuclear waste, perhaps—into the hole, in the opposite sense to its rotation. (That is, if the hole were rotating clockwise, as seen from its north pole, the spaceship would dump its load of waste into the horizon with a counterclockwise spin.) The waste would disappear forever into the hole, while the spaceship would acquire a kick of energy, leaving its orbit with more energy than with which it entered. The energy thus acquired can actually exceed the rest energy of the waste sent down the hole. From where did the extra energy originate? It came from the rotational energy of the black hole. Because material was sent into the black hole with opposite angular momentum, the hole is left with slightly less spin as a result of this encounter. In principle, a great deal of energy could be extracted from a Kerr black hole in this manner, but the amount of energy available is not infinite. As rotational energy is removed, the black hole must slow down. Eventually, all the rotational energy would be gone, and the Kerr black hole would become a Schwarzschild black hole. A classical Schwarzschild black hole is truly dead in the sense that no energy can be removed from it, not even by perturbing it.

Hawking radiation

Our discussion of black holes has so far dealt with the consequences of the classical theory of general relativity. When we try to add the strange world of quantum mechanics to that of general relativity, we find that it is not quite true that Schwarzschild black holes never lose energy. Recall that Einstein's equations imply that gravity itself possesses energy. In principle, then, the space-time curvature around a black hole could be tapped as a source of energy, even for the case of a nonrotating, stationary, Schwarzschild black hole. The Schwarzschild hole is dead only in the *classical* universe, that is, a universe without quantum mechanics. Although no one has succeeded in fitting gravity completely into a quantum mechanical description, it is possible to carry out quantum calculations on a background of a smooth, curved space-time (the semi-classical approach). Stephen Hawking found something remarkable from such a calculation: black holes are not completely black. They actually emit radiation, although the amount is extremely small for most black holes.

If nothing can escape from a black hole, what is the source of this **Hawking radiation**? Its existence depends upon the quantum mechanical effect known as the Heisenberg **uncertainty principle**, which states, among other things, that energy need not be strictly conserved for short times, provided that it is conserved, overall, for longer time

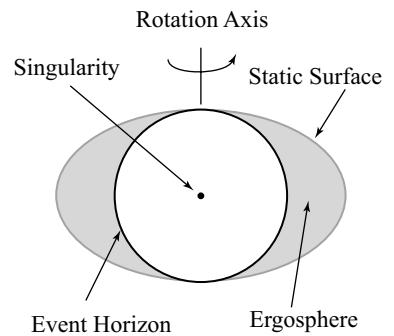


Fig. 9.4 Components of a rotating Kerr black hole. The ergosphere (shaded region) of a Kerr hole is located between the static surface (outer curve) and the event horizon (inner curve).

Black holes are not completely black

¹Greek *ergos*, work or energy.

The strange world of virtual particles

intervals. The greater the violation of energy conservation, that is, the more energy borrowed, the more quickly it must be repaid. In the quantum universe, even the purest vacuum is filled with a sea of **virtual particles** that appear as particle–antiparticle pairs, and then disappear in the fleeting interval of time permitted by quantum mechanics. Can one of these virtual particles ever become real? Yes, if that particle can acquire the energy to pay off its loan of energy before it comes due.

This effect can be demonstrated in the laboratory. Set up two parallel conducting metal plates separated by an empty gap. Onto these parallel plates place opposite electrical charges, creating a voltage difference and a strong electric field running from one plate to the other. Now place the apparatus in a vacuum chamber and increase the voltage across the plate. In the vacuum between the plates, negatively charged virtual electrons and positively charged virtual positrons are popping in and out of existence. But because they are doing so in the presence of an electric field, and because the field exerts a force upon charged particles that has opposite directions for opposite signs, a particle could be accelerated away from its anti-partner before they have a chance to annihilate and disappear. The virtual particles thus can become real, and we can measure this flow of electrons and positrons as a net electrical current from one plate to another. We see a current flowing in what we thought to be a vacuum! The energy for this process comes from the electric field, so energy is still conserved. Through a quantum mechanical process, the energy stored in the electric field is converted into matter, in accordance with Einstein's law $E = mc^2$. This phenomenon is called *vacuum breakdown*, and is an extreme example of a more general effect called *vacuum polarization*. The experiment demonstrates that the vacuum is not empty, but is filled with virtual particles and fields. The virtual particles can also affect real particles in very small, but measurable, ways.

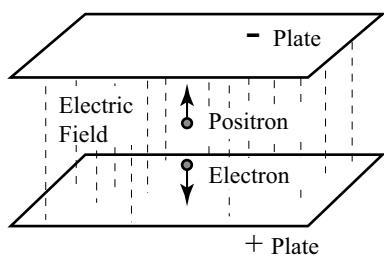


Fig. 9.5 Virtual particles can become real particles if they can tap into a source of energy. Here an intense electric field between two charged plates accelerates a virtual positron and a virtual electron away from one another, endowing them with the energy necessary to become real. The energy is obtained from the electric field.

Near a black hole, virtual pairs are created and destroyed, just as they are everywhere. But near the horizon, the tidal forces are strong, and stress from the tidal forces can be utilized to bring a pair of virtual particles into real existence. One member of this pair of particle Pinocchios falls into the horizon, while the other escapes to infinity. The emergent particles are Hawking radiation. The energy to create the particles comes from the energy of the gravitational field outside the hole. As that gravitational energy is lost to the creation of particles, the strength of the gravitational field is diminished, and the hole shrinks. Eventually, it *evaporates* and disappears from the universe. The final moments in its evaporation produce an intense burst of very high-energy particles and gamma rays. Since the radiation originates with virtual particle–antiparticle pairs, we should expect equal amounts of matter and antimatter to emerge from the hole. In fact, the easiest particle–antiparticle pairs to create are photons, particles of light. Photons, which are massless and are their own antiparticles, can appear at any energy level, whereas massive particles must be derived from at least as much energy as the sum of their rest energies. It is quite remarkable

that a black hole could be a source of *any* kind of particle, since the classical theory of relativity predicts that matter can only disappear beyond the event horizon. But now we find that, whatever the black hole originally consisted of, it emits Hawking radiation composed of photons, along with some matter and antimatter.

Most black holes are not very efficient at this process, however. Since the energy to create Hawking radiation comes from the tidal stresses, there must be substantial tidal stress present on the typical scale over which the virtual particles move. Because these particles exist only for a minuscule span of time, this scale is very small. The larger the hole, the weaker the tidal stress on a small length-scale; therefore, Hawking radiation is significant only for tiny holes. The time for evaporation of a black hole is proportional to the cube of its mass:

$$t_{\text{evap}} \approx 10^{10} \left(\frac{M}{10^{12} \text{kg}} \right)^3 \text{ yrs.} \quad (9.12)$$

The black hole decay rate

The wait for a solar-mass black hole to evaporate slowly due to Hawking radiation would be about 10^{65} years. This is considerably longer than the current age of the universe, roughly 10^{10} years. Nevertheless, if the future of the universe extends to infinite time, as the data indicate, then eventually all black holes will decay.

For Hawking radiation to be of much significance in the present universe, the hole must be a mini-hole. The only black holes that would be evaporating now would have masses of the order of 10^{12} kg, with a corresponding Schwarzschild radius of about 7×10^{-16} m; a mini-hole indeed. There is no observational evidence for the existence of such mini black holes. Whereas large black holes can form from the collapse of ordinary astronomical objects such as stars, the only conditions under which mini-holes could form would be inhomogeneities in the very early universe. There are severe constraints on the numbers of such tiny holes that could be produced in the big bang, making it doubtful that any such mini-holes exist. For a black hole of stellar or greater mass, Hawking radiation would be essentially undetectable, and would have no significant effect over most of the life expectancy of the universe. It is, however, a genuine phenomenon for even the largest black holes.

The fact that black holes radiate means that they have a *temperature*. An ideal emitter, or blackbody, radiates a continuum spectrum of photons, and that spectrum is uniquely determined by its temperature. The higher the temperature, the more energetic the spectrum. Remarkably, Hawking radiation turns out to be blackbody radiation. The temperature of this radiation, and thus of the black hole, is given by

$$T_{\text{bh}} \approx 10^{-7} \left(\frac{M_{\odot}}{M} \right). \quad (9.13)$$

Black hole temperature

The radiation emitted by a solar-mass black hole is very small, so it has a low temperature, only 10^{-7} K above absolute zero. Larger holes have even lower temperatures; a black hole with a mass of $10^6 M_{\odot}$ has a temperature of only 10^{-13} K.

The realization that black holes emit radiation led Hawking to a wonderful unifying concept for black holes. Following an idea of Jacob Bekenstein, Hawking had already developed a theory of the merging of two black holes; such a merger forms a single black hole, with a surface area that is larger than the combined surface areas of the previous two separate holes. This was the law of black hole areas: regardless of anything black holes might do, whether they collide, gain more matter, or add mass by any other means, the result will always be a hole with a larger surface area than it had before. This is very reminiscent of the second law of thermodynamics, which states that the entropy, that is, the disorder, of an isolated system must always increase. If the size of a black hole can be equated with its entropy, then that implies that the black hole should be described by the laws of thermodynamics, which in turn means that it *should* have a temperature. Hawking radiation accounts for that black hole temperature, allowing Hawking to formulate all these ideas into the laws of **black hole thermodynamics**. We shall return to this topic later, for it has tantalizing implications for the evolution of the universe.

Black hole exotica

White holes

Hawking radiation may seem quite odd, but there are even stranger things allowed by classical general relativity theory. One of these is the *white hole*, a kind of mirror image of the black hole. In a white hole, nothing can get *in*; it can only come *out*. Rather than space-time flowing into the horizon of a black hole at the speed of light, space-time flows out of the horizon of a white hole at the speed of light. Although this is intriguing, we know of no way in which a white hole could form. A black hole can be created in a straightforward manner, by the gravitational collapse of a material object such as a star. A white hole would have to be inserted into the universe as an initial condition. And where would it get the matter and energy that would pour forth? Just because it is difficult to imagine does not mean that it cannot be, but in the absence of any evidence to the contrary, we can say with some confidence that white holes do not exist.

Wormholes

The mathematics of the Schwarzschild solution admits the possibility of another odd beast, the *wormhole*. The space-time curvature produced by a Schwarzschild black hole can be envisioned as a kind of funnel in space-time. A wormhole is a connection from one such space-time funnel to another. Might the wormhole form a gateway from one point in space-time to another? Could a spaceship travel through wormholes to reach very distant locations in both space and time? Unfortunately, this is not the case, at least not for the Schwarzschild wormhole. All paths through the wormhole that avoid the singularity are spacelike; that is, they can be traversed only at speeds greater than that of light. As we have seen, a massive particle cannot travel such a worldline. Moreover, the full solution for Schwarzschild wormholes shows that they are dynamic and

evolve. They turn out to be unstable; they pinch off, trapping anything within them at the singularity. Clearly, this is not a desirable feature for a transportation system.

Kip Thorne and his colleagues have examined the wormhole solution in more detail; they found circumstances under which it might be possible to construct a wormhole with a route that could be followed by a timelike path. This most likely will not help us find shortcuts from one galaxy to another, however; such a wormhole requires exotic conditions that are unlikely to exist in the physical universe. Moreover, it is still unclear whether it might yet violate physical laws, and even if it does not, it is probably too narrow for anything but an elementary particle to follow. It seems, then, that wormholes may be useful as a plot device in science fiction, but have little, if any, relevance to the real universe. Why, then, do relativists study them? Aside from the intellectual pleasure of exploring such a unique topic, the study of wormholes and the possible quantum effects near them is a way of investigating the nature of quantum gravity. The odd properties of Hawking radiation, wormholes, white holes, and the like, provide insights to the properties of that as-yet undeveloped theory.

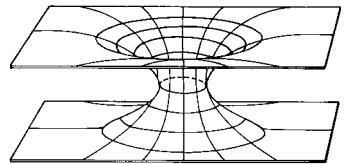


Fig. 9.6 A wormhole is a solution of Einstein's equations that appears to connect two different universes or different regions of one universe. However, the classical wormhole is unstable and cannot be traversed by a timelike worldline.

Black holes in the present universe

We have mentioned several exotic aspects of black holes, but always with the caveat that these effects are probably not important in the present universe. What, then, is the relevance of black holes now? Do they exist, and, if so, do they play any important roles, or are they merely mathematical oddities of the equations of general relativity? At the very least, their mere existence seems to be nearly certain. Black holes must be the end stage of the evolution of very massive stars. Upon consuming all its nuclear fuel, if a star is left with a core mass greater than the upper limit for a neutron star, collapse is inevitable. Black holes might also form at the center of dense clusters of stars, or in the cores of galaxies, perhaps as those galaxies are created. Because general relativity requires a black hole to form if the conditions are appropriate, the universe may well harbor countless black holes of varying sizes, ranging from modest black holes left behind after supernova explosions to huge monsters residing at the center of galaxies. Yet by their very nature, black holes are *black*; they emit no light, and the minuscule flux of Hawking radiation from any such moderate-sized black holes would be completely undetectable. How, then, can we see them?

The answer is that while we cannot see the holes, we can infer their presence by their effects on the light and matter that we *can* see. We have already mentioned how a passing black hole can distort the light from a distant star or galaxy in a way we might be able to detect. But any object with mass can produce a gravitational-lens effect, and it would be difficult to distinguish a lens created by a black hole from one created by a dim, but otherwise normal, star. The easiest way to

detect black holes is by their gravitational attraction on other matter. There are many possible interactions between a black hole and its surroundings, depending on the mass of the hole and the kind of matter in its vicinity. Evidence for the existence of compact sources of energy has been found for over thirty years. The energy can be liberated by a number of processes that might occur near the horizon. Stars and gas can fall into the gravitational well of a black hole. Stars can be torn apart by tidal forces; gas can be heated to enormously high temperatures, compressed, and shocked. Gas orbiting a black hole can be whipped around at extraordinarily high velocities in a very small region of space. Such phenomena now can be studied with ever increasing detail with a new generation of telescopes, both ground-based and space-based, such as the *Hubble Space Telescope* and the *Chandra* X-ray telescope.

Where might be the first place search for black holes? Since we believe that supermassive stars collapse to form black holes, there should be numerous stellar-mass black holes in our own galaxy. Finding them would not be easy because, as we have noted, an isolated black hole will produce no luminosity. Hence our first candidate locations should be binary stars, stellar systems consisting of two stars orbiting one another. In a binary system, the gravitational effect of a black hole will influence its visible companion in a detectable way. In some rare cases, we can observe the wiggles in the motion of a star with an unseen partner, and deduce the mass of the object from Kepler's laws. A number of systems are known in which the mass of the invisible companion must be greater than the upper limit for a neutron star.

Astronomers had not really given this type of system much thought until the early 1970s, when the X-ray satellite *Uhuru* detected powerful X-rays coming from the constellation Cygnus. This source, designated Cygnus X-1, proved to be a binary system that emitted energetic X-rays, but quite irregularly; the X-rays flickered over a very short time interval, about 0.01 second. Observations made with optical telescopes determined that the system included a hot, massive star. Wobbles in its motion made it possible to surmise that it has a companion with a mass of about five to ten times that of the Sun. Furthermore, this companion could not be detected by optical telescopes. The X-ray flickering is thought to occur in hot gas near the invisible companion. The rapidity of the variation is significant in establishing the size of the companion, because no object can vary in a systematic and regular fashion on timescales shorter than the time it takes light to cross it. As an analogy, imagine a huge marching band, spread out over such a large area that it takes 10 seconds for sound to travel from one end to the other. Now imagine that the musicians are blindfolded, and must play from what they hear. Such a band could not play staccato notes in unison every half second. The sound would arrive at a distant listener spread out over a 10-second interval. Since the speed of light is the fastest speed attainable, the largest region of an astronomical body that can be in causal contact over a time interval Δt is of size $c\Delta t$. In the case of the unseen member of the Cygnus X-1 binary, light can travel

Black holes in binary stars

only approximately 3000 km in 0.01 second. If the radius of the dark star is indeed 3000 km or less, it is a little smaller than the size of the Earth. Existing stellar theory cannot accommodate such a large mass for such a small star in any form other than a black hole. Neutron stars cannot have a mass more than about three times that of the Sun; thus the unseen companion is either a black hole, or something unknown to current theory.

The X-rays emanating from Cygnus X-1 are produced in gas that is lost from the normal star and drawn to the black hole. If two stars in a binary system are sufficiently close, gas from one star can be pulled away and fall onto the other star. If the other star is a compact object, such as a black hole, the gas falls deep into a strong gravitational field, becoming very hot and radiating high-energy photons before disappearing down the hole. Because the gas possesses some angular momentum, it orbits around the black hole, flattening into a spinning disk of gas called an **accretion disk**.² Turbulence in the disk causes the gas to spiral slowly toward the black hole. As the gas falls into the gravitational well of the black hole, it loses gravitational potential energy in exchange for a gain in other forms of energy; specifically, half of the gravitational energy is converted into heat energy. Collisions between the infalling gas and the matter already occupying the region close to the horizon could also compress and heat the gas. A sufficiently hot gas will emit X-rays, just as a cooler gas emits visible light. How much energy might be thus liberated when gas falls into a black hole? There is no clear answer to this question, as it depends upon details of the behavior of the infalling gas, but theoretical estimates range from a few percent, to as much as 40%, of the rest energy mc^2 of the gas. Considering that nuclear reactions release at most about one percent of the rest energy of the matter, it is clear that gravity power is a far more efficient means of generating energy. Thus, accretion disks can make black holes detectable from great distances, albeit only indirectly.

Accretion of gas onto a black hole can produce high-energy radiation

Some remarkable direct evidence for the existence of black holes has emerged recently: the detection of the gravitational redshift from gas near an event horizon. The first such observation was reported in 1995. A group of astronomers led by Y. Tanaka used an X-ray satellite to observe the core of a member of a class of galaxies called Seyfert galaxies; these galaxies are believed to harbor supermassive black holes at their centers. The astronomers found a severely redshifted X-ray emission line of iron. The redshift was consistent with light originating within a distance of approximately three to ten times the Schwarzschild radius of a black hole. Further observations of this type have found other examples, providing additional direct evidence for supermassive black holes in the centers of unusually luminous galaxies. At last, the black hole seems to have emerged from the pages of texts on general relativity, and shown itself to be as much a resident of the real universe as the stars.

²See also Chapter 5 and Figure 5.8.



Fig. 9.7 Illustration of a binary system consisting of a hot, massive star and a companion black hole. Gas is drawn from the normal star and forms an accretion disk around the black hole. (STScI/NASA.)

A supermassive black hole in the Milky Way

As astronomers gained the ability to observe in wavelength bands outside the visible, new phenomena were discovered that pointed to black holes and other compact objects as important inhabitants of the universe. In the early 1960s, when radio astronomy was yet a young science, astronomers mapped the sky at radio wavelengths, finding an abundance of radio sources. Most could be identified with known objects. One of these was the center of the Milky Way Galaxy, which is located in the constellation Sagittarius. We cannot see the center of our own Galaxy in optical wavelengths because of absorption by intervening clouds of cosmic dust, but it shines brightly in the radio. Radio telescopes have been used to map out the core of the Galaxy, revealing a complex region of swirling gas around an intensely radio-bright core known as Sagittarius A*, or, in abbreviated form, Sgr A*. Sgr A* is believed to lie at the very center of the Galaxy. But what is it? Radio observations show that the gas near Sgr A* is moving very rapidly. This is consistent with a significant gravitational field. Could the source be a supermassive black hole? Dust prevents us from studying the center in optical light, but some infrared light penetrates the obscuring clouds. Fortunately, near the center is a number of red giant stars that are both cool and bright, meaning that they emit significant fluxes of infrared radiation. These stars have provided a remarkable way to weigh Sgr A*. For over 10 years, powerful infrared telescopes have imaged the core of the Galaxy, locating the positions of these giant stars. From the data, the stars' orbits around Sgr A* can be traced. For some stars the orbital periods are short enough that the full orbit can be mapped. Kepler's laws can then be employed to compute the central mass, and the best fit to the data indicate that Sgr A* is a point mass of about $3 \times 10^6 M_{\odot}$, ruling out any possibility except a supermassive central black hole.

As it happens, our Galaxy's central black hole is relatively quiet. Although it produces radio and X-ray emission, the total brightness is low compared to the capabilities of supermassive black holes. For examples of highly luminous sources we must turn to observations of the cores of other galaxies. Many powerful radio sources are associated with optical galaxies. One of the brightest radio objects in the sky is Cygnus A, a

The discovery of quasars

galaxy that is 500 million lightyears distant. Cygnus A throws out over 10 million times as much radio energy as an ordinary galaxy. But when a number of bright radio sources were first discovered in the early 1960s, not all could be traced to optical galaxies; many such sources were at the time indistinguishable in optical wavelengths from stars. It was clear, however, that they could not be normal stars; normal stars are dim in the radio, and these were bright. More mysterious yet was the presence of unrecognizable emission lines in their optical spectra. Some scientists went so far as to propose that an unknown element existed in these objects.

The resolution to this puzzle came in 1963 when astronomer Maarten Schmidt recognized that these strange lines were the usual lines of hydrogen, but redshifted so much that they appeared in a completely unexpected portion of the spectrum. Redshift over cosmic distances indicates distance. The large redshifts that were measured for these objects implied fantastic distances, up to *billions* of lightyears. To be visible over such distances, the objects had to be almost unimaginably luminous. These objects outshine even the brightest galaxies by factors of 100 or more. In some cases, a luminosity 10,000 times greater than that of an ordinary galaxy would be required for an object at such a great distance to appear so bright. It was soon discovered that the light output of many of these objects fluctuates considerably over short intervals of time. The distance light travels over the interval of the variations sets an upper limit to the size of the source; changes in appearance over times of approximately a day or less means that the light must be coming from a region less than about a lightday in size. Hence a tremendous quantity of energy is pouring forth from a region comparable in extent to the solar system. Clearly these objects, whatever they were, were not stars. Since it was uncertain *what* they might be, and because on photographic plates they appeared as unresolved starlike points, they were called **quasi-stellar objects**, which is often shortened to **quasars**; they are also sometimes referenced by their acronym **QSO**.

The rapid oscillations and the starlike appearance point to a very compact energy source for quasars. Stars alone could never provide so much energy; what could? The best theory available requires supermassive black holes, holes with masses from several million up to a billion times that of the Sun. Such holes have Schwarzschild radii as large as the orbit of the Earth around the Sun, and would be surrounded by hot gas spiraling into the hole through a huge accretion disk. The whole system would be comparable in size to our solar system, and could process each year several solar masses' worth of gas. If a black hole could release just 10% of the rest energy of this gas via the accretion process, then the consumption of one solar mass of gas per year would provide enough energy for a luminosity roughly 100 times that of a garden-variety galaxy. A typical spiral galaxy might shine with the brightness of 10^{11} to 10^{12} Suns, while an average quasar emits 10^{13} to 10^{14} solar luminosities.

Quasars are not the only place where we might find supermassive black holes. The center of a normal galaxy represents another place

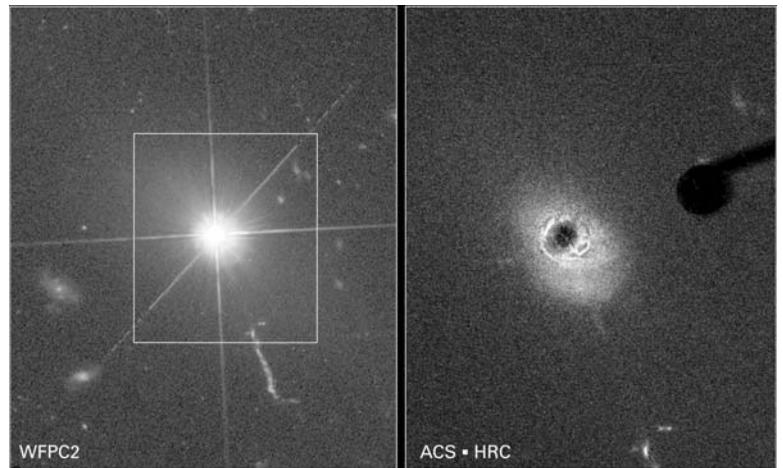


Fig. 9.8 *Hubble Space Telescope* photograph of a quasar. On the left, quasar 3C273 is an unresolved point source. On the right, light from the central quasar is blocked, revealing the surrounding host galaxy. This image shows a spiral plume, a dust lane and other complex structures. (NASA, A. Martel (JHU), H. Ford (JHU), M. Clampin (STScI), G. Hartig (STScI), G. Illingworth (UCO/Lick Observatory), the ACS Science Team and ESA.)

Black holes in active galactic nuclei

where gravitational collapse might occur; the larger the galaxy, the more prone to collapse its core might be. In fact, a small minority, about one percent, of the galaxies we observe are **active galaxies**; that is, they emit more than just ordinary starlight. Active galaxies occur in diverse shapes and sizes. One category is known as *Seyfert galaxies*; these appear to be typical spiral galaxies, but have abnormally bright centers with bright emission lines from hot gas. *BL Lacertae* objects are distant elliptical galaxies with a rapidly varying, unresolved point of nonstellar emission in their cores. *Radio galaxies* are giant elliptical galaxies that produce large amounts of radio energy in their central regions; examples include the aforementioned Cygnus A, and the famous galaxy M87 in the constellation Virgo.

The common feature of all these galaxies is that they emit copious amounts of nonstellar energy from a relatively small region in their centers, or nuclei. These bright cores are thus called **active galactic nuclei**, or **AGNs**. The AGNs that are brightest at radio wavelengths show an even more remarkable feature: *radio jets*, beams of radio-emitting matter, probably in the form of energetic gas, which appear to be shot from the very center of the galaxy. Some active galaxies possess two symmetric jets, directed oppositely away from the center. Sometimes only one jet is observed to emerge from the galaxy, but maps of the radio energy reveal *radio lobes*, large regions of diffuse radio emissions, on both sides of the galaxy. The single jet almost always runs from the center of the galaxy to one of the radio lobes; the partner jet on the other side of the galaxy is believed to be present, but unseen because it is beaming away from our line of sight. Some of the most powerful jets are gigantic, as much as three million lightyears in length. These huge jets have been powered over their lifetimes by enormous quantities of energy, as much as 10 million times the rest energy of all the matter in the Sun. Some jets are moving so fast that they vary over short timescales; anything that changes on a human timescale is astoundingly fast, by

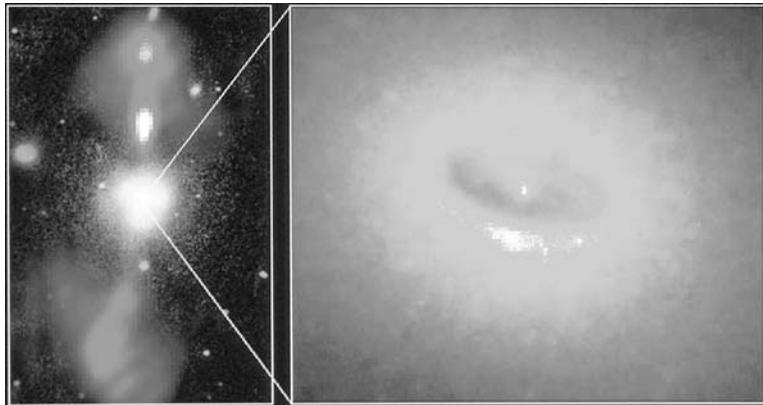


Fig. 9.9 *Hubble Space Telescope* observation of the core of an active galaxy, NGC 4261. On the left is a photograph combining data from ground-based optical and radio telescopes, showing powerful jets emanating from the core of the galaxy. The right-hand photo is an *HST* image of the central region of the galaxy, possibly showing an accretion disk. (H. Ford, Johns Hopkins; W. Jaffe, Leiden Observatory; STScI/NASA.)

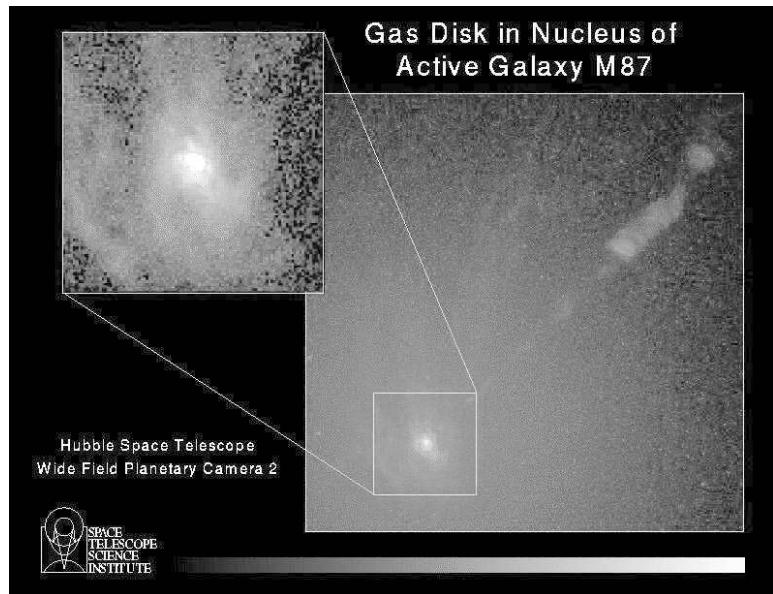
astronomical standards. Some jets appear to move faster than light, but this can be explained as an illusion caused by the beaming toward us of a jet whose gas is moving at speeds near that of light. We can thus infer that the most energetic jets consist of gas moving at relativistic speeds.

Jets require a compact energy source at the center of the galaxy. It must be a source that is capable of beaming huge quantities of energy in a specific direction for a very long time; it must also be capable of processing millions of solar masses of matter into energy over the jet's existence, at high efficiency. The best candidate for such a powerhouse is a supermassive black hole. Gas would be squirted out from an enormous accretion disk in the two directions perpendicular to the disk, along its axis of rotation. One potential power source for the jets might be the rotational energy of a Kerr black hole. A spinning black hole represents a huge reservoir of available energy. One possible means by which that energy might be extracted involves magnetic fields, generated in the surrounding accretion disk and connecting it with the black hole. As the field lines are wound up, they accelerate and focus the outflowing jets. This picture is somewhat speculative, but plausible. The study of active galactic nuclei, jets, accretion disks, and supermassive black holes is one of the most active areas of research in astronomy, both in theory and in observations. New observations continue to provide better data, by which theories can be tested, but also present us with new mysteries.

The giant elliptical galaxy M87 exhibits many of the phenomena associated with active galactic nuclei. It has been known for many years that this galaxy is special. It is unusually large, even for a giant elliptical; its volume is nearly as great as that of the entire Local Group. It sits at the apparent center of a very large cluster of galaxies, the Virgo Cluster. Its core is prodigiously energetic, and a well-defined jet shoots from the heart of the galaxy. This jet is bright not only at radio wavelengths, but at optical and higher wavelengths as well. The jet is observed to emit *synchrotron radiation*, a well-defined pattern of wavelengths characteristic of electrons spiraling around a magnetic field. It was long suspected that M87 might harbor a black hole, but it is some 50 million

The rotational energy of Kerr black holes could power jets and other galactic activity

Fig. 9.10 The right-hand image shows a jet emerging from the core of the giant elliptical galaxy M87. (Compare Figure 4.19.) The *Hubble Space Telescope* image on the left is a closeup of the core. Emission lines in the central core show systematic Doppler shifts, specifically blueshifts on one side of the core and redshifts on the other. The sharp peaks of observed spectral lines and their wide separation indicate a large velocity difference, which can be easily explained only by the presence at the center of the galaxy of a massive black hole, about which this gas is orbiting. (H. Ford, Z. Tsvetanov, A. Davidson, and G. Kriss, Johns Hopkins University; R. Bohlin and G. Hartig, STScI; R. Harms, L. Dresel, and A. K. Kochhar, ARC; Bruce Margon, University of Washington. STScI/NASA.)



lightyears distant, and even the best ground-based telescopes could not clearly resolve the motions in the innermost regions of the galaxy. Once the *Hubble Space Telescope* was repaired in 1993, M87 was one of its first assignments for spectroscopy. The results were spectacular; gas in the central 60 lightyears of M87 rotates much more quickly than can be easily explained by any hypothesis other than that it is orbiting around a supermassive black hole.

All these energetic objects have an apparently very compact energy source and an astonishing output of energy. Observations over the last few decades have shown that quasars are themselves examples of prodigiously active galaxies. They are so far away that we see them when the universe was much younger than it is now; studying them provides clues to the history of the universe. Quasars are often found in association with other objects that definitely have the appearance of galaxies, and which have similar redshifts. In a few cases, it has been possible to detect the faintest wisp of spirals around some QSOs themselves. If quasars/QSOs are indeed active galaxies, then they join the lineage as its most extreme members. If all are powered by black holes, they can be explained by differing scales: the larger the central black hole, and the more gas available for its appetite, the greater the energy it would generate. If this hypothesis is correct, active galaxies in general, and especially quasars, could not sustain such an outpouring of energy for long periods. We might see active galaxies during an explosive stage of their existences. They blaze for only a short time, on cosmological timescales, then, when the black hole has devoured all the matter readily available in its vicinity, the galaxy settles into quiescence. Perhaps it will sit placidly for the remaining lifetime of the universe, or perhaps

another source of gas will replenish the accretion disk and cause a new outburst. We can see only a snapshot of the universe as it is during our short lifespans; we cannot watch the evolution of an active galaxy. It is as if we could visit a family reunion only once, seeing various members of a family each at a single age; from that information, we could try to construct a hypothesis of how a given individual would age and change throughout life. We do much the same with active galaxies, making our best effort to understand them with the data available to us.

Chapter Summary

The first exact solution to the Einstein equations was found by Karl Schwarzschild. This solution, the Schwarzschild metric, describes the vacuum exterior to a sphere of mass M . The metric coefficients provide the mathematical description of gravitational time dilation and length contraction outside the sphere. The solution also introduces a new quantity, the Schwarzschild radius, $R_s = 2GM/c^2$. At the Schwarzschild radius the gravitational time dilation goes to infinity and lengths are contracted to zero. The black hole is a phenomenon that is predicted by the Schwarzschild metric; it is an object whose mass lies entirely within its Schwarzschild radius. The Schwarzschild radius marks an event horizon, a point of no return around a black hole. Once inside the event horizon, nothing, not even light, can escape. An observer at infinity watching a probe approach a black hole would see the probe's signals redshifted further and further, till at last the redshift would approach infinity and no more photons escaped. The distant observer would never see the probe cross the event horizon. To a sufficiently small probe, however, nothing unusual would occur at the horizon. But once across the horizon, the probe is doomed to fall into the singularity at the center.

There are other important radii near a black hole. The last stable orbit at $3R_s$ is the closest distance at which a material particle can orbit a black hole; any closer and the particle must move radially inward or outward. The photon sphere is defined by the radius $1.5R_s$ at which gravity bends the path of photons so much that light orbits the hole circularly. The significant bending of light by strong gravitational fields also leads to an effect known as a gravitational lens. Such a lens is produced when light passes through the gravitational field of a very massive object, such as a large galaxy, a cluster of galaxies, or a black hole. Lenses such as these provide important cosmolog-

ical data; in particular, images of and distances to very remote galaxies can be obtained.

Despite their extraordinary properties, black holes are actually quite simple. The no-hair theorem states that a static black hole is completely described by three quantities: mass, angular momentum, and charge. A black hole would be unlikely to retain any electrical charge for long in the real universe, but spinning black holes should be common. A black hole that has angular momentum is described by the Kerr metric. For a Kerr black hole the single event horizon of a Schwarzschild black hole splits into an ellipsoidal outer static surface and an inner event horizon. Between the two surfaces is the ergosphere. At the static surface, it is necessary to orbit at the speed of light opposite the rotation of the hole just in order to stay still; inside this point even light is compelled to orbit in the direction of the hole's spin.

Classical black holes are truly black, but quantum mechanics predicts that they can emit Hawking radiation. This is the emission of particles (mostly photons) from just above the event horizon. Hawking radiation is black-body in nature and permits a temperature to be assigned to a black hole. The temperature leads to a thermodynamics of black holes and to a definition of their entropy; black holes are found to have the highest entropy of any object in the universe. However, Hawking radiation is minuscule for all black holes of any significant size.

Black holes have other possible exotic properties. The complete black hole solution forms a wormhole, which might join two distant regions of space-time. Closed time-like paths may be associated with wormholes, although it is unlikely that anything larger than a subatomic particle could traverse these paths. Moreover, wormholes are dynamic and pinch off, making them useless for transportation, since anything in the wormhole would be crushed at the singularity. Another strange solution to the Einstein

equations is the “white hole,” which in many respects mirrors the black hole; rather than disappearing into the hole, matter appears at a white hole, and nothing can remain in the white hole. However, there is no evidence that white holes could form in the physical universe.

Black holes are believed to be produced in the final collapse of the most massive stars in the universe. If the hole

is located in a binary system it can accrete gas that can become very hot and emit powerful X-rays. Active galaxies are galaxies that have energetic, nonstellar activity in their cores. The best model for the central engine of active galaxies is a supermassive black hole surrounded by a huge disk of gas, an accretion disk, which slowly spirals into the hole and releases energy.

Key Term Definitions

black hole An object that is maximally gravitationally collapsed, and from which not even light can escape.

escape velocity The minimum velocity required to escape to infinity from the gravitational field of an object.

Schwarzschild radius The radius of the event horizon of a nonrotating black hole of mass M , equal to $2GM/c^2$.

event horizon A lightlike surface that divides space-time into two regions; that which can be observed, and that which cannot. The Schwarzschild radius of a nonrotating black hole is an event horizon.

singularity In classical general relativity, a location at which physical quantities such as density become infinite.

coordinate singularity A location at which a particular coordinate system fails, such as the Schwarzschild metric coordinates at the Schwarzschild radius of a black hole.

cosmic censorship The principle that singularities are never “naked,” that is, they do not occur unless surrounded by a shielding event horizon.

quantum gravity A unification of gravity and quantum field theory, not yet achieved.

Birkhoff’s theorem A theorem of general relativity which states that all spherical gravitational fields, whether from a star or from a black hole, are indistinguishable at large distances. A consequence of this is that purely radial changes in a spherical star do not affect its external gravitational field.

photon sphere The radius around a black hole at which light paths are gravitationally bent into a circle, thus causing the photons to orbit the hole.

gravitational lens A massive object that causes light to bend and focus due to its general-relativistic effect upon the space-time near it.

no-hair theorem The theorem that the gravitational field of a black hole is entirely determined by only its mass, angular momentum, and any electric charge.

Kerr metric The metric that describes the space-time around a rotating black hole.

static surface The surface surrounding a Kerr black hole at which even light cannot resist being dragged along in the direction of the rotation of the hole.

ergosphere The region of a rotating Kerr black hole between the static surface and the event horizon.

Hawking radiation Emission of particles, mostly photons, near the event horizon of black holes due to the quantum creation of particles from the gravitational energy of the black hole.

uncertainty principle The principle of quantum mechanics which states that the values of both members of certain pairs of variables, such as position and momentum, or energy and time interval, cannot be determined simultaneously to arbitrary precision. For example, the more precisely the momentum of a particle is measured, the less determined is its position. The uncertainty in the values of energy and time interval permits the quantum creation of virtual particles from the vacuum.

virtual particles Particles that exist only as permitted by the uncertainty principle.

black hole thermodynamics The theory that permits a temperature and an entropy to be defined for black holes.

accretion disk A disk of gas that accumulates around a center of gravitational attraction, such as a white dwarf, neutron star, or black hole. As the gas spirals inward, it becomes hot and emits light or even X-radiation.

quasar (QSO) An object that emits an extremely large luminosity from a small region. Invariably found

only at large redshifts and hence distances. Also called a *quasi-stellar object* or *QSO*.

active galaxy A galaxy whose energy output is anomalously high. About 1% of galaxies are active. Most contain an AGN at their cores.

active galactic nucleus (AGN) An unusually bright galactic nucleus whose light is not due to starlight.

Review Questions

- (9.1) A neutron star is very compact and dense, but it is not a black hole. If a typical neutron star has a mass of $2.5M_{\odot}$, what is its Schwarzschild radius? If the actual radius of the neutron star is 30 km, how does this compare to the Schwarzschild radius?
- (9.2) Excited atoms of hydrogen emit light with a wavelength of 1216×10^{-6} m (that is, 1216 Ångstroms). Suppose that you detect this line in emissions coming from a very compact source within the Milky Way Galaxy, but you measure its wavelength to be 1824×10^{-6} m. What might account for the change in wavelength? If the light originated from near a black hole, from how close to the Schwarzschild radius, expressed as a fraction of R_s , was the radiation emitted?
- (9.3) Define and distinguish *singularity*, *coordinate singularity*, and *event horizon*.
- (9.4) You are the commander of an exploratory mission to a black hole. You launch a robotic probe on a trajectory that will take it into the black hole. The probe has an internal clock and sends a radio pulse back to your ship at a fixed interval, in the reference frame of the probe. What effects do you observe in the signals from the probe as it approaches the black hole? What might you observe if the hole is rotating?
- (9.5) You plan to take a spaceship to the photon sphere and hover above the black hole to observe the back of your head. What sort of acceleration will you experience as you hover at this point? (Answer qualitatively, e.g., small, comparable to one g , several times g , much bigger than g , incredibly huge.)
- (9.6) If the Sun were to collapse and form a black hole, how would the orbit of the Earth be affected? Would any gravitational radiation be produced?
- (9.7) You observe an X-ray source to vary on a timescale of 0.001 seconds. What is the upper limit for the size of the X-ray emitting region? What is the mass of a black hole with a Schwarzschild radius of this size?
- (9.8) The Earth is a rotating body whose gravity is not as strong as that of a black hole. Does the Earth exhibit any of the effects we discussed for black holes, such as gravitational redshifts, frame dragging, or gravitational time dilation? How are the Earth and a black hole alike, and how are they different?
- (9.9) For a rotating (Kerr) black hole, define and distinguish the *static surface*, the *event horizon*, and the *ergosphere*. How might energy be extracted from a Kerr black hole? Is this an infinite source of energy for some advanced civilization?
- (9.10) How massive would a black hole have to be in order for it to evaporate due to Hawking radiation in only one year? How big is that mass compared to some object with which you are familiar? (On the surface of the Earth, $1\text{ kg} \approx 2.2\text{ lb}$.)
- (9.11) Discuss the leading model for X-ray emissions from a binary system which might include a black hole. If nothing can escape from inside a black hole, from where is the energy coming?
- (9.12) Discuss the unified theory for active galactic nuclei. Include topics such as the possible identity of the central engine, the origin of jets and radio lobes, and the range of activity.

- (9.13) The Milky Way Galaxy may have a total mass of around $10^{12} M_{\odot}$, or one trillion times the mass of the Sun. What is the Schwarzschild radius for the Milky Way? Divide the mass of the galaxy by the volume of such a black hole ($\frac{4}{3}\pi R_s^3$) to obtain the

density of such a black hole. (The mass of the Sun $M_{\odot} = 2 \times 10^{30}$ kg, and the Schwarzschild radius of the Sun is 3×10^3 m.) How does that density compare with water, which has a density of 10^3 kg per cubic meter?

Part IV

The Big Bang

This page intentionally left blank

The Expanding Universe

10

The history of astronomy is a history
of receding horizons.

Edwin Hubble

For at least as long as written history has existed, humanity has set its sights upon understanding the shape, scope, and history of the universe. To this task we bring our senses, our experience, and our reason. This was as true for ancient cosmologists as it is for modern scientists. Today, however, our senses are augmented by powerful tools, we benefit from the accumulated and recorded experience of many generations, and we have developed mathematical languages that provide an efficient means to systematize our reasoning. In this chapter, we shall see how these advances led to one of the greatest discoveries in history: the expanding universe.

The biologist or the geologist is accustomed to gathering data in the field; the chemist, to the direct manipulation of molecules in the test tube; the physicist, to the construction of apparatus to measure a particular phenomenon. The astronomer, in contrast, must be content to look. The only exceptions are the occasional meteorite, and the several hundred kilograms of Moon rocks returned to Earth by the Apollo astronauts. And although we have sent robotic investigators to other planets to do our experiments and sampling remotely, these efforts are confined to our own solar system, and will remain so for the indefinite future. Almost everything we know about the universe, at scales larger than that of our solar system, comes from the electromagnetic radiation we collect. Even what knowledge we might obtain about the nonluminous contents of the universe must be inferred from its effect upon the matter we can see. Despite this fundamental limitation, the astronomer can learn a great deal through careful observations of the light that reaches the telescope.

We tend to think of light as just something that illuminates our surroundings, but there is much more to it than our eyes can see. Light, or electromagnetic radiation, can be found in a full range of energies. This distribution of energies makes up the electromagnetic spectrum. What we call visible light is just one small range of light's energy spectrum. Complete knowledge of an astronomical object requires observing across the full spectrum. As a possibly more familiar example, consider the spectrum of sound frequencies heard while listening to music; different notes of music correspond approximately to different frequencies

Key Terms:

- **redshift**
- **blueshift**
- **cosmological redshift**
- **nebula**
- **nova**
- **supernova**
- **cosmic distance ladder**
- **parallax**
- **parsec**
- **luminosity distance**
- **standard candle**
- **extinction**
- **Cepheid variable**
- **Hubble constant**
- **Hubble law**
- **Tully–Fisher relation**
- **Hubble expansion**
- **peculiar velocity**
- **cosmological constant**
- **dark energy**
- **de Sitter model**
- **scale factor**
- **comoving coordinates**
- **Robertson–Walker metric**
- **curvature constant**
- **cosmic time**
- **Hubble time**
- **Hubble length**
- **Hubble sphere**
- **lookback time**

The science of astronomy depends on observations

Astronomers study the full spectrum of electromagnetic radiation

of sound wave. Everyone would probably agree that a Beethoven symphony is best appreciated by registering the arrival of individual notes and distinguishing among the different sound frequencies. There would be much less benefit in summing the total sound energy of all the notes arriving at a microphone during the performance! Similarly, while it can be quite useful to measure the total luminous flux from an astronomical object, a far greater wealth of information is obtained by performing *spectroscopy*, the measurement of the quantity of light energy at each wavelength or frequency. Spectroscopy was developed and applied to astronomy late in the 19th century. The significant advances that made this possible were the recognition that different elements have unique spectral signatures; the development of photographic techniques that not only could make a permanent record of a star's spectrum, but also were much more sensitive than the eye; and finally, the construction of substantially larger telescopes that made it possible to collect enough light to perform spectroscopy on faint objects.

So much useful information is derived from the spectrum of electromagnetic radiation that telescopes devote most of their time to spectroscopy. In analyzing a spectrum, an astronomer considers many issues. Which lines are present? Are they emission or absorption lines? Are they shifted from their laboratory-based standard positions? What is the overall distribution of energy in the spectrum? The particular lines present in the spectrum of an astronomical object can identify the composition of the emitter, while other characteristics of the spectrum give clues to the temperature of the object, its internal motions, and the processes occurring within it. From this we learn such things as how the light was emitted, which elements are present and in what abundances, what are the velocities of the gas that emitted the light, and what population of stars a galaxy contains.

When an astronomer compares the spectrum of a star or a galaxy to laboratory standards, the emission or absorption lines associated with individual molecules and elements are typically not located at exactly the same wavelengths as the standards. Since we now have a good understanding of the elements, we would not be inclined to hypothesize the existence of new elements. The discrepancy, instead, is explained by an overall shift of the spectrum. A shift to longer wavelengths and lower energies is called a **redshift**, while a shift to shorter wavelengths and higher energies is a **blueshift**. These are the generic terms for these shifts, even if the radiation detected is not near the red or blue part of the visible spectrum. Because the relative *spacing* between the lines of a given element never changes, and a redshift or blueshift occurs for the spectrum as a whole, it is still possible to identify elements in a shifted spectrum by comparison to measurements in our Earthly laboratories.

How might the spectrum of light be blue- or redshifted? The most mundane, and prevalent, process is the ordinary Doppler effect, which is a consequence of the relative motion of the source of the light and our detector. Doppler shifts are easily detected for nearby objects, and are an important source of information about motions in the universe. The

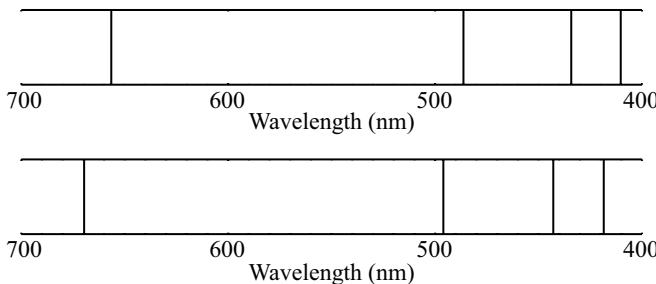


Fig. 10.1 Top: a spectrum of the element hydrogen, showing four emission lines from the Balmer sequence. Bottom: the same four lines shifted toward the red by an amount corresponding to a redshift of 2%.

formula for the nonrelativistic Doppler shift is

$$z = \frac{\lambda_{\text{rec}} - \lambda_{\text{em}}}{\lambda_{\text{em}}} = \frac{v}{c}. \quad (10.1)$$

The shift is symbolized by z ; this quantity is also often called the redshift regardless of whether it actually represents a redshift or a blueshift. The shift z is a positive quantity when light is redshifted, that is, when λ_{rec} is greater than λ_{em} , and is negative when it is blueshifted. This equation must be suitably modified when dealing with relativistic velocities, as we showed in our discussion of special relativity.

Even though the Doppler shift almost always tells us only about the relative *radial* motion of the source¹, it still provides abundant information. For nearby objects, an overall Doppler shift indicates whether they are approaching or receding from the Earth. Many objects show both a redshift and a blueshift, sometimes superposed upon an overall shift; this indicates that part of the object is approaching and part receding, perhaps relative to a bulk motion of the object as a whole. Such spectra reveal that the object is rotating, and can even enable astronomers to measure its rotation rate. In a few cases, the spectrum of a star is found to shift back and forth at a regular interval, indicating that the star is in orbit around another object whose light cannot be resolved over the glare of its brighter companion. Sometimes careful searches for the partners of these *spectroscopic binaries* fail to find the companion; such a system may be a candidate to contain a neutron star or even a black hole. The Doppler shift is without doubt one of the most important measurements in astronomy.

Another source of spectral shifting is the gravitational redshift. Gravitational redshifts occur when photons climb out of a strong gravitational field to a point where the field is weaker; conversely, gravitational blueshifts occur when light falls from a weaker to a stronger point in the field. The gravitational field of the Earth is extremely weak in comparison with the fields of stars or galaxies. Since the photons from those objects were emitted from a much stronger gravitational field than that of the Earth, any gravitational shifts we would observe from astronomical sources would be redshifts. Such gravitational redshifts are

The redshift is an essential astronomical measurement

Both ordinary motion and gravity can produce redshifts

¹The exception is the transverse Doppler shift due to relativistic time dilation. See Chapter 7 for more details.

A new effect: the cosmological redshift

almost always extremely tiny and difficult to measure, unless they originate from compact objects such as neutron stars or white dwarfs. The Schwarzschild gravitational redshift provides an example of how space-time curvature, as described by a metric function, can affect light as it moves through space-time. Other solutions to Einstein's equations must also have the potential to produce redshifts and blueshifts. In this chapter we will introduce the **cosmological redshift**, which is produced by the overall metric, that is, the space-time geometry, of the universe. The cosmological redshift is a consequence of the fact that the universe is not static and stationary, but is a dynamic, changing space-time: an *expanding* universe.

The discovery of the external universe

The Milky Way as a disk

Today we are accustomed to thinking of the Milky Way Galaxy as merely one among billions of galaxies in the universe, and not a particularly significant galaxy at that, except, perhaps, to us. But before the 20th century, few imagined that external galaxies might exist. One of the first original thinkers to grasp this idea was Thomas Wright, who published in 1750 the suggestion that the Earth and Sun lay within an enormous shell of stars. He pointed out that our view through this shell would appear as the Milky Way, the band of light that circumscribes the heavens. Wright's book stimulated the philosopher Immanuel Kant to modify and extend the hypothesis. Kant realized that the appearance of the Milky Way in the sky could be explained if it was shaped like a disk, with our Sun somewhere within the disk. In 1755 he published a book describing a universe inhabited by a finite Milky Way surrounded by many similar Milky Ways, all clustering in groups of ever-increasing size. However, such a prescient view of the universe was a decidedly minority position at the time. In the late 18th century, little was known about even the Milky Way; evidence for other galaxies was entirely lacking. Astronomers had catalogued many fuzzy patches of light, called **nebulae** from the Latin word for cloud, but the true nature of these objects was unknown. It was suspected that the nebulae were blobs of glowing gas spread among distant stars. Such gas clouds do exist in space. They contain mostly hydrogen, and glow with light from atoms energized by hot stars embedded within the nebula. Most such clouds that are easily observed are very near the solar system; a well-known example, visible to the naked eye in moderately dark skies, is the Orion Nebula, the faint patch in the sword below Orion's belt. Not all nebulae, however, are so obviously gas clouds.

The first of the great catalogs of nebulae was compiled in 1780 by Charles Messier, primarily as an aid to astronomers searching for comets. Though many more nebulae have been discovered since Messier's time, the objects described in his catalog are still known today as the Messier objects, and are designated by the letter "M" followed by a number; an example is the great elliptical galaxy M87 in the constellation Virgo.

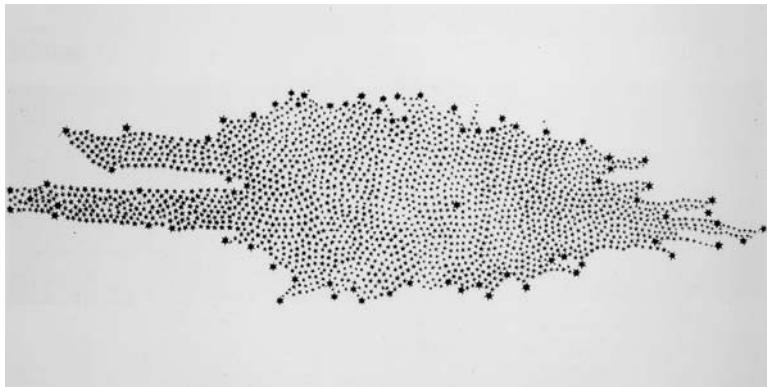


Fig. 10.2 Herschel's depiction of the distribution of stars in the Milky Way. The Sun was thought to be near the center of a somewhat irregular distribution of stars. (Courtesy of Yerkes Observatory.)

The list of nebulae was soon expanded by William Herschel, a professional musician turned astronomer who almost single-handedly created the field of galactic astronomy. Herschel set about a detailed study of the distribution of stars in the Milky Way, using telescopes of his own design and construction. In 1785 Herschel published the first diagram of the Milky Way, which he called a "detached nebula." He suggested that many of the mysterious nebulae could be similar agglomerations of stars. As telescopes improved, some nebulae revealed an overall structure. The first direct evidence in support of Kant's view of the universe came from observations carried out by Lord Rosse (William Parsons) in 1845. Rosse observed that some nebulae had a distinct spiral structure, suggesting to him that they could be "island universes" similar to our own Milky Way. The nebulae that resembled whirlpools of light were designated *spiral nebulae*. Astronomers remained divided over the nature of these spiral objects; some agreed with Rosse and Kant that they were external galaxies, while others believed them to be spiral conglomerations of stars within the Milky Way, as globular clusters are spherical associations of stars within and around our Galaxy. Some argued that the whirling appearance of the spiral nebulae suggested that these were new stars and solar systems, caught in the early stages of formation.

The difficulty in elucidating the nature of the spiral nebulae was compounded by the lack of a good determination of the size of the Milky Way, and our location within it. Herschel had described the Milky Way as a somewhat small, amorphous disk of stars, with the Sun near the center. What was not appreciated at the time was that dust within the Milky Way blocks our view through the Galaxy itself. In particular, this effect led Herschel to underestimate considerably the size of the Galaxy. Obscuration by dust also made the spiral nebulae appear to be preferentially located out of the plane of the Milky Way, suggesting to 19th-century astronomers that the distribution of these nebulae was somehow related to the Galaxy, and hence that they must be associated with it. On the other hand, there was some evidence that individual stars were present in the spirals; if so, their faintness would argue for a great distance, well beyond the boundaries of the Milky Way. With

The mysterious "nebulae"

How large is the Milky Way?

no knowledge of the actual size of the Milky Way, it was difficult to determine whether the nebulae were located inside or outside our Galaxy. If the Milky Way was as large as some believed, the small apparent diameters of the spirals meant that they would be fantastically remote, if they were comparable to the Milky Way in size. And if that were the case, then no individual stars could possibly be seen if they were only as bright as known stars in the Milky Way.

One piece of evidence came from the observation of novae in some of the nebulae. A **nova**, from the Latin *nova stella* or “new star,” is an abrupt increase in the brightness of a star, due to an enormous flare-up. Novae can reach a maximum luminosity that is as much as 100,000 times the Sun. Since they do not represent the end of a star, but rather a large and temporary increase in its energy output, they are fairly common and can even repeat themselves. This is in contrast to the much brighter, but rarer, **supernova**, which does result from the destructive explosion of a star.² Although it can be seen to much greater distances, the observation of a supernova is chancy, and was especially so before the days of systematic searches for them. An additional problem was that astronomers at the time did not realize that such a thing as a supernova existed. In 1885 astronomers observed what they thought to be a nova in the Andromeda nebula; although it did not appear to be unusually bright, it was nevertheless comparable in luminosity to the rest of the nebula. Clearly, it seemed to them, the nebula could be neither too far away, nor composed of billions of unresolved stars. Ironically, what had been observed was actually a supernova, which *does* rival the brightness of an entire galaxy. In fact, the relative faintness of the supernova was evidence for a substantial distance. But with no distance reference, and no understanding of the distinction between a nova and a supernova, astronomers assumed that what they had seen was an ordinary nova.³ Such preconceptions can confuse a scientific question for years, as the history of astronomy vividly illustrates. The 19th century passed with no resolution of these issues in sight.

During the first twenty years of the 20th century, the nature of the spiral nebulae remained one of the major scientific controversies. New and important data were introduced in 1912 by Vesto Slipher, who measured the spectral shifts, and hence the radial velocities, of some of the spiral nebulae. He found that many of them had velocities much greater than is typical for stars within the Milky Way. In fact, some of them had velocities that might be so great as to exceed the escape velocity from the Milky Way, a finding which certainly argued in favor of the island universe model. Other data seemed to contradict this, however. Among the more influential observations of the time were those of Adriaan van Maanen of the Mount Wilson Observatory, who claimed in 1916 to have directly observed rotational motion in the spiral nebula M100. If visible

The island universe model was the hypothesis that the spiral nebulae are external galaxies

²See Chapter 5 for a discussion of the properties of novae and supernovae.

³It was Hubble’s independent determination of the distance to the Andromeda Galaxy in 1925 that established the existence of supernovae as a new and distinct phenomenon.



Fig. 10.3 Harlow Shapley (1885–1972). Shapley's measurements of globular clusters enabled him to determine the size of the Milky Way and the location of the Sun within it. (Courtesy of Yerkes Observatory.)

transverse motion could be observed in only a few years' time, then the nebula could not possibly be very far away, else the implied rotational speed would be in excess of the speed of light. Although it was not realized at the time, van Maanen's observations were simply erroneous. Acceptance of his results, however, led many astronomers to consider them the final blow against the island-universe hypothesis. Even today, it is unclear how van Maanen, a highly competent and experienced astronomer, could have committed such a gross error. Perhaps his interpretation of his data was affected by his beliefs. The subject began to become clearer in 1917, when Heber Curtis found three faint novae in spiral nebulae. Based on this, he correctly rejected the Andromeda nova as anomalous, and employed the dimmer novae to conclude that the spiral nebulae must be millions of lightyears away.

At the same time that van Maanen and Curtis were carrying out their research on the nebulae, a young astronomer named Harlow Shapley, also working at the Mount Wilson Observatory near Los Angeles, set about to make a careful study of the size and extent of the Milky Way. He focused his attention on globular clusters, the gigantic, spheroidal agglomerations of stars that orbit the Milky Way. Shapley's work on the

Shapley charts the Milky Way

spatial distribution of the globular clusters showed that they occupied a roughly spherical region. He postulated, correctly, that the center of the sphere was the center of the Milky Way—and the Sun was nowhere near it. Before Shapley attacked the problem, the Milky Way was believed to have a diameter of fifteen to twenty thousand lightyears, with the Sun at its center. Shapley concluded that its true diameter was nearly 300,000 lightyears, with the Sun located near the edge. Unfortunately, various errors led him slightly astray; although the Milky Way is indeed much larger than anyone dreamed at the beginning of the 20th century, Shapley's estimate of its diameter was too large by roughly a factor of three. His somewhat erroneous conclusion led him to the belief that the Milky Way was so enormous and grand that the spiral nebulae must be mere satellites about it, spiral counterparts of the globulars.

These issues crystalized in 1920, when a formal debate was held on the subject at the National Academy of Sciences in Washington, D.C. Representing the local hypothesis was Harlow Shapley; the island universe hypothesis was championed by Heber Curtis. In all fairness, Shapley was mainly concerned with establishing the size of the Milky Way. Yet it was felt that his success would defeat the island universe hypothesis *en passant*, since the distances required for the spiral nebulae would simply be too unimaginably great, if the Milky Way were as large as Shapley believed. Although Curtis, an experienced public speaker, mounted a more focused argument during the 1920 debate and was, in the formal sense, generally judged the winner, Shapley proved more persuasive in the larger scientific discussion, on the basis of his masterful calculations of the size of the Milky Way. Interestingly, Shapley's determination of the diameter of the Milky Way, although an overestimate by a factor of three, was nevertheless fairly accurate, in comparison with the small diameter fashionable at the time; yet Curtis' view on the nebulae ultimately proved correct. In truth, they were both right; the fact was that nobody was then quite ready to conceive of the true vastness of the cosmic distance scale.

The Shapley–Curtis debate

The cosmic distance ladder

How would an astronomer have gone about measuring distances in 1920? It is a difficult problem, as it remains today, because the methods available to measure very large distances are not necessarily highly accurate. The procedure that is used has come to be known as the **cosmic distance ladder**, because each successive distance scale depends on accurate measurements at the earlier stages (or rungs) of the process. The first of these rungs is the most direct and accurate method, but it is useful only for nearby stars. This is the method of **parallax**, which is the measurement of the apparent shift of a star's location on the celestial sphere due to the motion of the Earth in its orbit. Parallax was sought in vain by early astronomers, but the stars are too distant for these shifts to be detected with the naked eye. Today we can measure

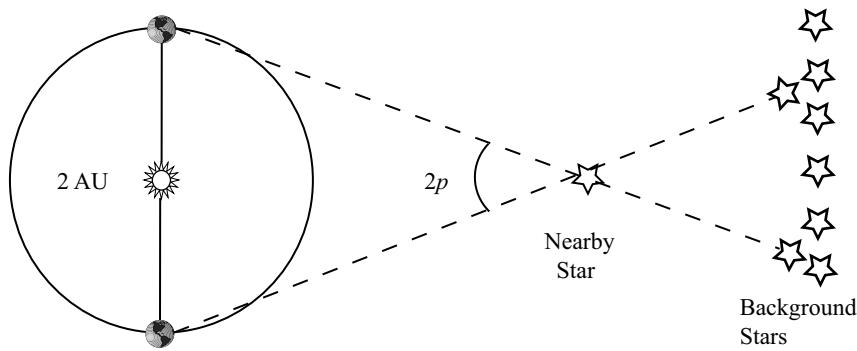


Fig. 10.4 The geometry of parallax measurements, showing the angular shift in a nearby star's position due to Earth's orbit. The figure is enormously exaggerated. The largest parallax angle observed for any star is less than 1 arcsecond.

these shifts, and once the very small corresponding angle has been determined, triangulation is used to compute the distance. The method is very similar to that employed by a surveyor to establish the distance to a mountaintop by measuring its angle from two different positions. Since the resulting triangles are, in the astronomical case, extremely long and thin, we may use the formula

$$d = \frac{2\text{AU}}{2p} = \frac{\text{AU}}{p}, \quad (10.2)$$

where an AU (astronomical unit) is the mean distance of the Earth from the Sun, p is the measured parallax angle, and d is the desired distance to the star. (Notice that the parallax angle is defined in terms of half the total baseline, as illustrated by Figure 10.4.) If d is to be determined in absolute units, such as meters, p must be expressed in radians. However, p is usually measured in seconds of arc, $1/3600$ th of a degree. From this we derive the unit of length called the **parsec**, which is that distance producing one second of arc of parallax over the baseline of the Earth's orbit. A parsec (pc) corresponds to 3.26 lightyears.

The parallax angles of even fairly nearby objects are incredibly tiny, and the distances of the stars remained beyond the reach of astronomers until technological improvements in telescopes and their mountings made it possible to determine star positions with great accuracy. After centuries of futile attempts by many observers, Friedrich Wilhelm Bessel announced in 1838 that he had measured the parallax of the inconspicuous star 61 Cygni. Without the aid of photography, Bessel found a parallax angle of 0.3 seconds of arc, from which he computed a distance for this star of 10.9 light years, a figure very close to the modern result of 11 light years. Parallax is the only *direct* method of determining interstellar distances; it requires no knowledge of the structure or brightness of the star, nor does any intervening matter affect the result. Parallax measurements demand extremely accurate determinations of an object's position at different times of the year. For very distant stars, the blurring of the star's image by the optical distortion that is inevitable in any telescope swamps the minute shifts in its apparent position. This problem can be overcome by observing from space; in 1989 the European Space Agency launched the *Hipparcos* satellite to measure parallax

Celestial distances are obtained from a hierarchy of methods building from direct parallax measurements

angles. The satellite provided an accuracy approximately 10 times that of Earth-based observations; it measured distances for roughly a million stars to a distance of about 200 pc. Future missions, such as NASA's proposed *Space Interferometry Mission*, may be able to measure parallaxes to stars throughout the Galaxy. However, until such space missions are launched, we must turn to more indirect means to determine distances beyond our immediate stellar neighborhood.

The most common indirect approach is the method of **luminosity distance**. This approach depends upon the fact that if the intrinsic luminosity of an object is known, then a measurement of its *apparent brightness* makes it possible to deduce the distance traveled by the light. The amount of light received at a particular location from the source is reduced with increasing distance, because the energy emitted must spread out over a larger and larger sphere as it travels outward into space. (Figure 4.15 illustrates the geometry involved.) The surface area of a sphere is given by $A_s = 4\pi r^2$; hence the brightness, which is the energy per unit time per unit area, must diminish by the inverse square of the distance from the source. The apparent brightness received at the surface of the Earth from a star at a distance d_L is thus simply $L/4\pi d_L^2$; in this expression, L is the star's *luminosity*, its total output of energy per unit time. We may rearrange this formula to obtain

$$d_L = \sqrt{\frac{L}{4\pi b}}, \quad (10.3)$$

where b is the apparent brightness obtained from the light falling upon the telescope.

As a specific example, consider the energy radiated into space by the Sun. The Sun has an intrinsic luminosity of $L_\odot \approx 4 \times 10^{26}$ watts. Suppose an astronomer on a distant planet observes the Sun with a telescope whose effective mirror radius a is 2 meters, in Earthly units. The surface area of the mirror is thus $\pi a^2 = 4\pi \text{ m}^2$. Suppose further that the alien astronomer's telescope collects from the Sun an energy per unit time of 4×10^{-8} watts. The apparent brightness measured at this distance is thus $10^{-8}/\pi \approx 3 \times 10^{-9}$ watts per meter squared. Assuming the astronomer knew, or could estimate, the luminosity of stars of the Sun's type, the astronomer could then apply equation (10.3) to find that the distance of the Sun from the alien planet is 1×10^{17} meters, or approximately 10.6 lightyears. As suggested by this example, the quantity of energy received from even a nearby star is very small. As enormous as the total energy output of a star may seem, this energy spreads out through the vastness of space; only a tiny fraction arrives at the Earth and impinges upon a human retina, or upon the focus of a telescope. In comparison, a typical flashlight emits about 5 watts of energy from a lens of approximately 15 cm^2 , for a brightness of roughly 333 watts m^{-2} . It should be obvious why astronomy demands huge telescope apertures, sensitive detectors, and dark skies.

A major weakness of the method of luminosity distance is its reliance upon a knowledge of the absolute luminosity of the target object. The

Luminosity distances are based on the inverse square law of light

first step in obtaining absolute luminosities depends on determining distances to nearby star clusters, particularly the Hyades cluster in Taurus, by other, more direct, means, such as parallax, or by determinations of the inherent proper motions of the stars of the cluster. By measuring the apparent luminosities of the stars in the same cluster, we can then work backwards, using our knowledge of distance and apparent luminosities to compute the intrinsic luminosities of the different kinds of stars found in the cluster. Once the luminosity of a particular type of star is known, it becomes a **standard candle**, a term referring to any object of known luminosity. As it turns out, stars of a given mass and age vary little in luminosity; if we find another star of the same type that is too far away for its parallax to be measured, we can, in principle, obtain its luminosity distance. Unfortunately, a further set of difficulties complicates this type of measurement. One particularly important confounding effect is the presence of intervening dust, which reduces the apparent brightness of a standard candle beyond that explained by distance alone. This phenomenon is known as **extinction**, and it was a significant source of systematic error in the 1920s, when the existence of interstellar dust was not yet recognized. This phenomenon specifically led Shapley to overestimate the size of the Milky Way, and confused the study of spiral nebulae for several decades.

In the early debates over the nature of the spiral nebulae, the greatest problem was that ordinary stars were the only well-established standard candles; but in the nebulae, individual stars were simply too faint to detect and too small to resolve. Even today, it is difficult to observe individual stars in galaxies, and it was nearly impossible with the technology available in the 1920s. If it is difficult to see a star, it is even more difficult to determine its type, since accurate spectra are required, and spectroscopy demands the collection of quite a lot of light. The further the galaxy, the more severe this problem becomes. Heber Curtis attempted to use novae as a standard candle in the Andromeda Galaxy, but the presence of a supernova complicated the issue. Beyond that, novae themselves are not precisely consistent in their luminosities. A new and better standard candle was required.

Astronomers require bright standard candles to obtain distances to galaxies

The Hubble law

Edwin Hubble, one of the astronomers most instrumental in changing our view of the universe, now enters the story. In the 1920s, Hubble undertook a systematic survey of spiral nebulae. The critical breakthrough occurred in 1924, when Hubble detected a star of the type called a **Cepheid variable** in the Andromeda Nebula, an object that is now known as the Andromeda Galaxy. Cepheid variable stars vary in brightness with a fixed periodicity. The periodicity differs from star to star, over a range of about 3 to 50 days, but Henrietta Leavitt discovered in 1912 that the periodicity is a function of the star's maximum luminosity. Specifically, the brighter the star, the longer the period. Herein lies

Cepheid variables are an important standard candle

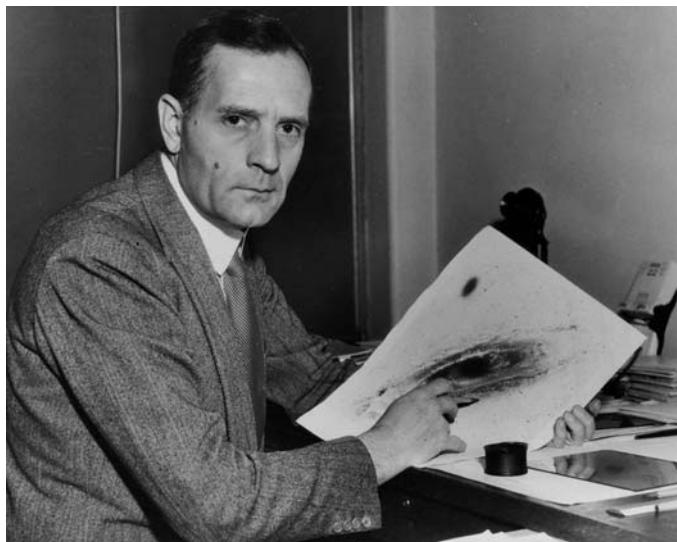


Fig. 10.5 Edwin Hubble (1889–1953). Best known for his discovery of the expanding universe, Hubble also was the first to measure the distance to the Andromeda galaxy, proving that the so-called “spiral nebulae” were external galaxies. (Hale Observatories, courtesy AIP Emilio Segrè Visual Archives.)

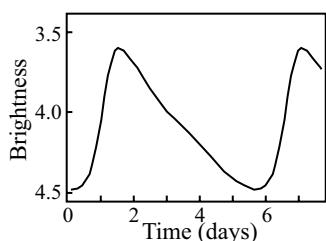


Fig. 10.6 The apparent brightness of a Cepheid variable changes with time. By observing the light curve, its period of variation can easily be determined. The longer the period, the more luminous the Cepheid.

the importance of Cepheid variables. By measuring the star’s apparent brightness along with its period of variation, a relatively straightforward operation, Hubble was able to compute the Cepheid’s luminosity, and hence the distance to the nebula in which it resided. Hubble’s discovery that the Andromeda nebula was very remote, far beyond the reaches of the Milky Way, settled the debate once and for all; the spiral nebulae were external galaxies. We now know that the distance to the Andromeda Galaxy is approximately 2 million lightyears (700 kpc). Andromeda is the nearest large galaxy to the Milky Way, to which it is mutually gravitationally bound.⁴ In modern astronomical usage, external star systems are always called galaxies; the term nebula is reserved exclusively for those objects that truly are clouds of gas and dust. Some texts continued to refer to galaxies as nebulae until well into the 1950s, however.

It is useful to reflect upon the significance of the discovery of other galaxies. This revelation was yet another blow to humanity’s anthropocentric cosmological point of view. First Copernicus removed the Earth from the center of the universe. Centuries elapsed during which the heliocentric theory was grudgingly accepted, but humans stubbornly retained their sense of specialness by shifting the center of the universe to the *Sun*. Again, appearances conspired to make it seem so: the band of the Milky Way has a nearly uniform brightness all around the sky, implying that we are at its center. Astronomers did not realize that dust obscured their view of the true center of the Galaxy. Then, with his study of globular clusters, Harlow Shapley proved that the Sun is not at the center of the Galaxy. But the question remained whether or

⁴The Andromeda Galaxy, also known as Messier 31 (M31), can be seen by the naked eye from a dark location. It is literally “as far as the eye can see.”

not the Milky Way constituted the bulk of the matter in the universe. Finally, Hubble showed that the Milky Way is not the only galaxy in the universe. We now know that the Milky Way is a good-sized, although typical, spiral galaxy, and a member of a rather insignificant group of galaxies falling toward a much larger cluster of galaxies. And even that attracting cluster is not one of the biggest of all clusters. The universe is filled with numberless galaxies, organized into huge structures stretching over millions of parsecs. In humbling humanity, 20th-century astronomy outdid even Copernicus.

By establishing their true nature, Hubble created a new branch of astronomy, the study of galaxies. Hubble was a pioneer on this new frontier. Using the recently completed 100-inch telescope on Mt. Wilson, Hubble developed a classification scheme for galaxies, based upon their appearance, which is still widely used today. Galaxies are grouped into two morphological classes, spiral and elliptical. Spiral galaxies have a disk-like shape with a central bulge; the disk contains spiral arms of greater or lesser prominence. Elliptical galaxies are, as their name suggests, ellipsoidal conglomerations of stars. They exhibit little or no substructure such as spiral arms or flattened disks. Hubble subdivided these groups further by developing classification criteria for galaxies of each type, based on details of their overall structure. The ellipticals were grouped on the basis of their overall ellipticity, from the nearly round E0 type, to the highly elliptical E7. Spirals were divided into two groups, those with prominent stellar bars extending across the nuclear region (the barred spirals) and those without. These groups were further classified by the tightness of the spiral arms and the compactness of the nuclear bulge. Hubble also created another category, the irregular galaxies, for galaxies whose appearance is, as the name implies, irregular. The Large and Small Magellanic Clouds are two nearby dwarf irregular galaxies.

As valuable as these discoveries were, Hubble's preeminent contribution to cosmology lay elsewhere. At the time of the Shapley–Curtis debate, it was known that the majority of the spiral nebulae showed redshifted spectra; by 1922 Vesto Slipher had found that the spectra of 36 out of a sample of 41 of these objects were redshifted. (The Andromeda Galaxy is an example of a galaxy that is blueshifted; it and the Milky Way orbit one another, and currently Andromeda is approaching us.) Following his triumph with the Andromeda Galaxy, Hubble, aided by Milton Humason, obtained dozens of galactic spectra with the Mt. Wilson 100-inch telescope. The galactic spectra provided redshifts, but redshifts alone gave little information, other than that the galaxies were nearly all receding. Hubble and Humason went further, and combined their redshift data with their measured distances to the galaxies.

Redshifts are easily measured; obtaining distances is the challenging part. Hubble began with Cepheids, but they quickly become too dim to function as standard candles for more distant galaxies. Beyond the Cepheid limit, other calibrations must be used. Hubble and Humason used the best standards known to them, beginning with the brightest

The Milky Way is just one galaxy in a universe of galaxies

Hubble's classification scheme for galaxies

Spectroscopy of galaxies

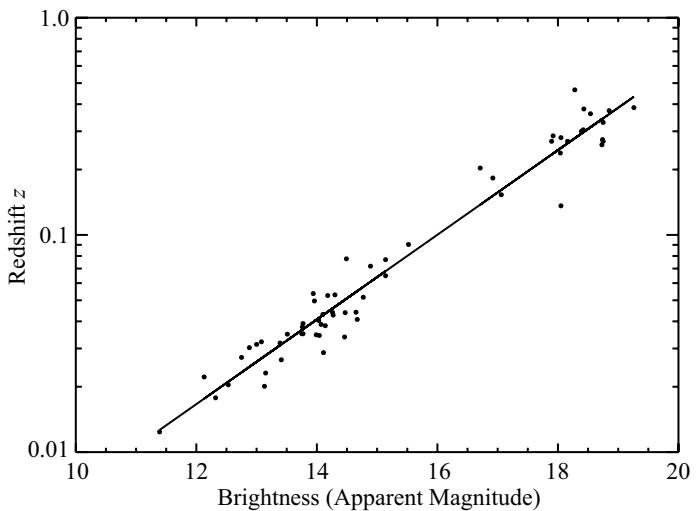


Fig. 10.7 A Hubble diagram obtained by plotting redshift against brightness (apparent magnitude) for bright galaxies in distant galaxy clusters. By assuming brightest cluster members are of comparable intrinsic brightness, the apparent brightness becomes a measure of distance.

supergiant stars of a galaxy. Stars cannot be infinitely bright, so there must be a cutoff; for most large galaxies, the brightest stars seem to be of similar luminosity. Hubble and Humason thus assumed that the brightest stars in any galaxy all have approximately the same luminosity. This hypothesis was checked using galaxies whose distances were found via Cepheids, and it worked reasonably well. The greatest difficulty with this approach is that the brightest *object* in a galaxy need not be its brightest normal *star*; for many galaxies, small clouds of extremely hot hydrogen gas may be the brightest point-like object. At cosmological distances, it can be very difficult to distinguish such emission regions from a star. Finally, at great enough distances, even the brightest stars fade into the general glow of the galaxy, and other techniques must be brought to bear. At very large distances, Hubble and Humason were reduced to using the apparent luminosity of the galaxy as a whole. They knew that the intrinsic luminosity of galaxies probably varied a good deal, but they hoped to limit the variation by comparing galaxies of the same Hubble classification. Despite the many potential sources of error, Hubble and Humason found that when redshift was plotted versus distance, the points were not randomly scattered about, but lay very close to a single straight line. In Hubble's own words, he had found a "roughly linear relationship between velocities and distances."

In fact, what Hubble had found directly was a linear relationship between *redshift* and distance ℓ , symbolized $z \propto \ell$. The distinction is subtle, but significant. Hubble measured the shifts in the spectra. A redshift can be caused by several factors, the most obvious of which is radial motion. In the cosmological case, however, the redshift is caused by the relativistic expansion of the universe itself, but this was not understood at the time that Hubble and Humason were compiling their data. Hubble interpreted the extragalactic redshifts in terms of the familiar Doppler shift. As it happens, the cosmological redshift and the

The Hubble law: redshift is proportional to distance

Doppler redshift behave the same way for nearby galaxies and small redshifts, up to approximately $z \leq 0.1$. For small redshifts, then, one may employ equation (10.1), in which the redshift is directly proportional to a velocity. Under these conditions, the graph of velocity versus distance will also be a straight line, and the redshift can be equated to a recession velocity. A straight line through the origin, that is, zero relative velocity at zero distance, implies a relationship of the form $v = H\ell$, where H , the slope of the line, is now called the **Hubble constant**. The general relationship

$$v = H\ell, \quad (10.4)$$

is called the **Hubble law**. The Hubble law is a *theoretical* description of the behavior of the universe. Hubble's observed redshift-distance relation provides the experimental evidence for this law. The Hubble constant in equation (10.4) must be determined by observation. It is usually expressed in units of kilometers per second per megaparsec. For example, if $H = 100 \text{ km s}^{-1} \text{ Mpc}^{-1}$, then a galaxy with a recessional velocity of 3000 km s^{-1} , corresponding to a measured redshift of 0.01 ($z = v/c$), would be located at a distance of 30 Mpc from the Earth. If $H = 50 \text{ km s}^{-1} \text{ Mpc}^{-1}$, then the same galaxy, with the same redshift and inferred recessional velocity, would be 60 Mpc distant.

The theoretical Hubble law

The current measured values for H fall mostly between 60 and 80 $\text{km s}^{-1} \text{ Mpc}^{-1}$, the range representing the different uncertainties in the specific techniques used to obtain distances. The best current value, consistent with a wide body of data from a variety of measurements, is $72 \text{ km s}^{-1} \text{ Mpc}^{-1}$. Hubble's original value for this same constant was around $500 \text{ km s}^{-1} \text{ Mpc}^{-1}$, a result drastically different from today's value. The reason for this discrepancy is instructive, for it illustrates the role of *systematic errors* in distance determination. Hubble did not realize that there are actually two classes of Cepheid variable stars; this was not discovered until 1952. Hubble was actually observing in external galaxies a brighter class of Cepheid, a type of star now designated Type I Cepheids; these are the classical Cepheids, like the prototype star δ Cephei, and they are intrinsically very bright stars. However, the period-luminosity relationship used by Hubble had inadvertently been calibrated for distance using another kind of variable star, now called Type II Cepheids, which are dimmer than Type I Cepheids by approximately a factor of four. As a result of this confusion, as well as other factors, such as failure to take extinction sufficiently into account, Hubble systematically underestimated distances to the Type I Cepheids by more than a factor of two. This is an example of a systematic error that can throw off a set of measurements by a much larger factor than that indicated by formal experimental uncertainties; stated uncertainties generally only account for the random errors that are inevitable in experiments. Because of systematic errors, Hubble believed that galaxies were, on the average, much smaller than the Milky Way. This should have been a clue that something was wrong, if we adhere to the Copernican principle. When more correct distances are employed

Systematic errors can be much greater than formal error estimates indicate

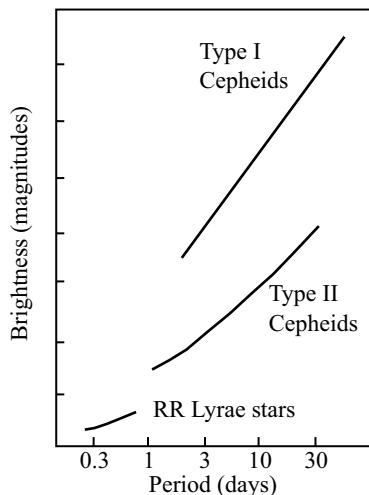


Fig. 10.8 The period–luminosity relation for variable stars. The period of variation increases with luminosity. There are two populations of Cepheids, Type I and Type II. Type II Cepheids are not as luminous and are therefore of less importance to cosmology.

Cepheids in the Virgo galaxy cluster were a HST key project

The Tully–Fisher relation is an indirect measure of distance to spiral galaxies

in our size estimates, we find that the Milky Way is a typical, perhaps even smallish, spiral.

Today, technology has improved the longest distance measurements. One of the most important recent developments was the launch of the *Hubble Space Telescope (HST)*. Above the distortions caused by the Earth’s atmosphere, the *HST* can resolve individual stars to a much greater distance than can most ground-based telescopes. One of the *HST*’s key projects was to find Cepheids in a sample of galaxies within 20 Mpc; the derived distances were used to calibrate additional distance indicators to obtain the Hubble constant. For example, *HST* detected Cepheids in several members of the Virgo cluster of galaxies. Once these Cepheids were found, it became possible to obtain a distance of 17 Mpc to the cluster.

The Virgo Cluster is the nearest large cluster, and detailed observations of it have been carried out for many decades. In particular, supernovae have been detected in the cluster and their light curves measured. This has led to the establishment of a specific type of supernova, called a Type Ia, as a standard candle. A Type Ia supernova results from the explosion of a white dwarf that was near the Chandrasekhar mass at the time of its demise. These supernovae have a very high peak brightness that shows relatively little variation from one Type Ia to another. It was thus long believed that these supernovae had the potential to form a good standard candle, due to the uniformity of the progenitors and the light curves. However, because none had been observed in galaxies for which an independent and accurate distance was known, the fundamental brightness calibration was lacking. This changed when the Hubble Cepheid data for the Virgo Cluster were collected. With the improved data came the realization that the peak of the Type Ia light curve correlates with the rate of decline in the light. Slow decliners are intrinsically brighter, thus enhancing their role as a standard candle. Supernovae are far brighter than Cepheids, making them visible at cosmological distances. More than any other single recent discovery, this has revolutionized cosmological observations.

Obviously, supernovae can be used only for galaxies in which one is observed. Other techniques must be employed for other galaxies. One of the most reliable methods for obtaining distances to spiral galaxies is the **Tully–Fisher relation**, named for its discoverers R. Brent Tully and J. Richard Fisher. The Tully–Fisher relation is a correlation between the width of one particular emission line of hydrogen, that at a wavelength of 21 centimeters in the rest frame of the hydrogen, and the luminosity of the spiral galaxy from which the emissions are observed. The main reason that there is a width at all to the 21-cm line is the rotation of the galaxy. Some of the gas is approaching, while some is receding, but at the resolution of ordinary radio observations, only the collective photons from all the gas in the galaxy are detected, causing the line to smear into a band. The broader the band, the faster the rotation of the galaxy and the greater its luminosity. The Tully–Fisher relation is based upon observations of many galaxies, not upon theory; however,

there is a simple, qualitative explanation that accounts for it well. A brighter galaxy obviously has more mass in stars than does a dimmer galaxy; we would thus expect the overall mass to be greater as well, including any nonluminous matter that might be present. If we assume that the ratio of total mass to luminous mass is roughly constant for most spiral galaxies, then the brightness of any spiral galaxy's emissions should be related to its total mass. From Kepler's third law, we know that a more massive galaxy would rotate faster. The width of the 21-cm line indicates the rotation rate of the galaxy; the relationship between total mass and brightness implies that the 21-cm line can stand as an approximate proxy for the total luminosity. The most remarkable aspect of this relationship is how good it is. It has proven to be one of the most useful distance indicators for very remote galaxies, though even it has its limitations; beyond approximately 200 Mpc, the width of the line becomes difficult to measure accurately.

Why is it so important to obtain distances to these remote galaxies just to determine the Hubble constant? If the only galactic motions were solely due to the **Hubble expansion**, then the measurement of one galaxy's distance would suffice. However, much of the scatter in a typical Hubble diagram is not due only to errors in distance. This is because the observed redshift is a composite of red- or blueshifts due to all velocities and gravitational effects. Most galaxies are members of clusters, and interact gravitationally with other galaxies. Gravity is a long-range force, so even the clusters can be influenced by other clusters. For instance, the Milky Way is primarily in a mutual orbit with the Andromeda Galaxy, but also, along with other members of the Local Group, orbits the Virgo Cluster. The intrinsic motion of an object due to its particular responses to forces such as local gravitational attractions is called the *peculiar motion* of the object; its velocity due to such movement is called its **peculiar velocity**. The term "peculiar" refers to the unique or particular velocity that a galaxy has in addition to its Hubble motion. The peculiar velocity is simply the velocity that results in the ordinary, classical Doppler shift due to the unique motions of a given galaxy, as distinct from the overall Hubble effect. The net redshift is a superposition of the peculiar Doppler shift upon the cosmological redshift.

For nearby objects, it can be very difficult to extricate the Hubble law from this combination of redshifts. However, because peculiar velocities will tend to be in all directions, both toward and away from us, their effect on the Hubble diagram will average to zero if we use a large number of galaxies at a given distance. Furthermore, peculiar velocities will all be less than some certain maximum amount, perhaps a few hundred kilometers per second, whereas the Hubble effect increases with distance. It follows that the relative importance of peculiar velocity is itself a function of distance. The closest objects, such as the other members of the Local Group, are dominated by their peculiar motions, and are essentially unaffected by the Hubble law. A little further out, matters become quite complicated. For example, peculiar motions play

A galaxy's motion independent of the Hubble expansion is its peculiar velocity

Distinguishing the Hubble law from peculiar velocities

a significant role in attempts to determine the Hubble constant from observations of the Virgo Cluster. The Virgo Cluster is, in cosmological terms, quite near. It is easiest to measure the distances of the closest objects accurately, such as by finding Cepheid variables in a galaxy. If the galaxy M100, a member of the Virgo Cluster whose distance has recently been determined to very high accuracy by the *HST*, showed only a cosmological redshift, then we could immediately determine the Hubble constant, since redshifts can be measured to very, very high precision; this is especially true for nearby objects whose spectra can be easily obtained. Unfortunately, the galaxies of the Virgo Cluster are executing complicated internal motions. More than that, the cluster is also a center of attraction for the Milky Way and its companions, increasing further the Doppler shift due to relative peculiar motions. All of this creates Doppler redshifts that are not insignificant compared to the cosmological redshift. Redshifts measured for the galaxies of the Virgo Cluster thus do not reveal an unambiguous cosmological redshift. In order to obtain a value for the Hubble constant from the measurement of the distance to M100, a model for the motions of the galaxies of the Virgo Cluster, and of the motion of the Milky Way toward them, must be employed to interpret the data. The distance to M100 serves mainly to establish a rung on the cosmic distance ladder so that more distant galaxies may be used in the determination of the Hubble constant.

For distant objects, those well beyond the Virgo Cluster, the cosmological redshift is so large as to completely swamp any peculiar velocities. The systematic increase of redshift with distance is the strongest argument that the cosmological redshift is truly cosmological. If it were due to peculiar motions of the galaxies, the redshift would show no tendency whatsoever; indeed, it would be expected that approximately as many distant galaxies would show blueshifts as redshifts, which is emphatically not observed. Some have argued that the large redshifts of quasars are due to peculiar velocities, mostly by appeal to a few anomalous cases that could easily be misleading. However, nearly all cosmologists agree that the data present overwhelming support for the interpretation that the major contribution to the redshifts of distant objects is the cosmological redshift, due not to any peculiar motions but to the *expansion of space itself*. This is a new and challenging concept. Usually, when we speak of a Doppler shift's implying a certain recession velocity, we mean that the shift is due to the inherent motion of the source relative to the receiver. But regarding the cosmological redshifts in such a manner could lead to a picture of galaxies streaming away from us. Such a picture implicitly places the Milky Way in the center of some great explosion, a point of view that is quite clearly inconsistent with the cosmological principle. From the Copernican principle, we should expect that we are not at the center of *anything*, much less some universal cataclysm. This means that the recession velocity in the Hubble law is very different from the kind of velocity to which we are accustomed. The cosmological redshift is due to the properties of *space itself*. Since we observe that all galaxies that are not gravitationally bound to the

The implications of the Hubble law

Milky Way show a cosmological redshift, and never a blueshift, they must be receding from us. What this observation implies is that space is expanding everywhere. Every galaxy sees every other galaxy expanding away from it. The overall motion of galaxies away from one another, due to the general expansion of the universe, is the Hubble expansion. In the next section we shall consider how such a strange notion arises from Einstein's theory of general relativity.

The theoretical discovery of a dynamic universe

In the first decades of the 20th century the astronomers, mostly in the United States, were enlarging the Milky Way, discovering external galaxies, and collecting the first hints of an overall cosmological redshift. Meanwhile the theorists, mostly in Europe, considered cosmology to be too speculative, almost metaphysical, and thus hardly worthy of serious scientific contemplation. But now and then a physicist or an astrophysicist dabbled in cosmology. Albert Einstein was among the first to investigate cosmology from a firm theoretical basis. Shortly after Einstein had completed the correct formulation of the equations of general relativity, with his usual scientific audacity he turned his attention to their implications for the entire universe. The new theory of gravity seemed to have properties that could solve some of the age-old questions of the universe. By admitting the possibility of space-time curvature, it was at last possible to construct a universe that was comfortably finite, thus avoiding the disturbing prospect of infinite space, yet without invoking an equally unfathomable edge. All that was necessary was to insert enough matter-energy into the universe to force it to curve space back upon itself, forming a spherical geometry that was both homogeneous and isotropic. Such a universe has a pleasing Machian property about it: the overall distribution of matter exactly determines the shape and size of space.

Einstein found that constructing such a model was easier imagined than done. The difficulty was that even in its relativistic form, gravity remains an attractive force. The tendency for a distribution of mass to undergo gravitational collapse, a problem that plagued Newton's clockwork universe, is not alleviated by confining the universe to a finite spherical domain; quite the opposite, relativity enhances the propensity of matter-energy to collapse. Einstein's equations predicted that his spherical universe, left to its own devices, would come crashing down on itself in about the amount of time it would take light to complete one circuit through the universe. Of course, a model in which the universe almost instantly collapsed upon itself was not very satisfying. It was, however, possible to adopt an alternative picture in which the universe was expanding, rather than contracting; the expansion would tend to counteract the pull of gravity. The important point was that, due to the omnipresent gravity, the universe could not remain still. Much like

Einstein was quick to apply general relativity to cosmology

Gravity causes the collapse of Einstein's spherical universe

a ball thrown into the air, it had to move. But this was still 1917, a time when, to astronomers, the universe consisted entirely of the Milky Way. Astronomers of the time may not have known too much, but this they did know: the Milky Way was not contracting or expanding. Stars were moving about, but there seemed to be no systematic expansion or contraction.⁵

Einstein was faced with a quandary. His equations predicted a dynamic universe, not the static universe that everyone believed to exist. Hence he decided, probably reluctantly, that the equations must be wrong. But how could they be wrong when they had worked so well for the corrections to the orbit of Mercury? The only possible answer was that there must be a term that is important only on the cosmic scale. How could such a term be accommodated without destroying the mathematical properties that had taken so long to establish? In formulating the Einstein equations, there is a point at which, essentially, an integral is performed. It is always possible, when integrating, to add a constant term; Einstein had initially set that constant of integration to zero, consistent with the notion that the force of gravity drops toward zero at large distances. But what if that constant were not zero? There could exist a term that is immeasurably small on the scale of the solar system, but which nevertheless creates a repulsive force at large scales, a force just sufficient to counteract the attraction of the matter in the universe. Such a force would have interesting properties. It would be zero on small scales, but would *increase* as a function of distance. Consequently, any nonzero (positive) cosmic force, no matter how small, would ultimately dominate on the largest scales.

Einstein's cosmic repulsion force, the cosmological constant

Einstein called his new constant, designated by the Greek letter Λ , the **cosmological constant**. It may seem at first glance to be a merely *ad hoc* term, but it enters the equations of general relativity as a perfectly legitimate possible contribution to large-scale gravity. Looking back at Einstein's equations (8.5), we see that there are two ways in which to regard Λ . The equations of general relativity state that the geometry of space-time equals mass-energy. The Λ term could be a constant of integration added to the geometry on the left-hand side of the equation; that is, it might be regarded as a mathematical correction. However, it could also arise from physical phenomena; whereas geometry, lurking on the left-hand side of the equation, is well defined in terms of established mathematical quantities such as the metric, the right-hand side is much less well established and could conceivably include new effects, or additional properties of matter and energy that are still unknown. The cosmological constant can thus also be interpreted as a contributor to the mass-energy of the universe that produces a repulsive force, a kind of negative energy term. In this case, Λ would correspond to a negative

⁵In fact, there is a systematic motion: the galaxy rotates. But at the time, the Milky Way was believed to be a somewhat amorphous disk of stars, with the Sun at the center. Coincidentally, 1917 was also the year that Shapley published his observations of globular clusters, indicating that the Sun was actually at the outskirts of the Milky Way, not the center.

energy associated with the vacuum of space itself. Astronomers now refer to this as **dark energy**. The cosmological constant will re-emerge later in this guise, not only in the context of the inflationary model of the early universe, but in new data about the structure and fate of our current cosmos.

Einstein introduced Λ to his equations to provide a balance between the attractive force of gravity and the repulsive force of the dark energy. Unfortunately, as it seemed at the time, the effort failed. It was not any artificiality of the cosmological constant that ultimately proved fatal to Einstein's static universe. Instead, the model proved to be *unstable*; while it could be set up in a static equilibrium, it simply did not remain static. In the static model, the balance between gravity and the repulsive force is delicate. The Λ force increases linearly with increasing distance, while ordinary gravity diminishes as the inverse square of the distance. Thus, if such a balanced universe were to expand just a bit, the repulsive force would grow while the attractive gravitational force would decrease. But this would mean that the universe would expand even more, leading to further decreases in gravity and more expansion. Conversely, if the universe contracted a little, the force of gravity would increase, overcoming the cosmological repulsion. The situation is analogous to that of a marble sitting on the top of an inverted bowl. The marble may be at rest (static), but any slight perturbation would cause it to roll away from the top. Similarly, the static universe was destined to move. Happily, the changing view of the universe eliminated the need for such a model. Hubble's data indicated that the universe, by then realized to be much bigger than the Milky Way alone, was, in fact, *expanding*. With Hubble's observational evidence of the reality of the expansion of the universe, Einstein attempted to retract his cosmological constant, going so far as to call it his "greatest blunder." This is probably too harsh, although one can certainly sympathize with his chagrin. Had he believed what his equations, in their original form, were telling him, he could have predicted the expansion of the universe before it was observed. Theory and observation are often at odds, and theory should remain mutable in the face of the experimental results. However, the problem here lay with the perennial assumption that humanity's observations up to a particular point in history provide a sufficiently representative sample of the universe from which to draw cosmological conclusions; this dates back at least to the first person who gazed across a wide open field, and concluded that the Earth was flat.⁶

The cosmological constant can be used to create many interesting alternative models of the universe, the simplest of which was published by Willem de Sitter in 1917, the same year as Einstein's static universe.

Einstein's static model does not remain stable

Hubble's observations imply an expanding universe

⁶It must be admitted that Einstein certainly had demonstrated the hubris to put his theory ahead of observation. When asked what he would think if the 1919 eclipse expedition had not observed the predicted shift of the stellar positions, he reportedly said "I would have had to pity our dear Lord. The theory is correct." In this case Einstein believed more in his theory's beauty, simplicity, and unity than in the ability of astronomers to measure the positions of stars accurately.

In contrast to the Einstein model, the **de Sitter model** contains no matter, just space-time and the repulsive Λ . In this model the geometry of space is flat, but the repulsive force causes space to expand exponentially. If a few stars were sprinkled into this otherwise-empty universe, they would recede from one another with velocities proportional to the distance. Given the obvious limitations of the de Sitter model, it is not surprising that it was not taken very seriously. But Slipher was contemporaneously publishing redshift observations and, for a while, the overall recession of the galaxies was even called the “de Sitter effect.” The clues were emerging, but the synthesis had not yet occurred. Only after the passage of another decade would the realization dawn that the universe is expanding.

A metric for an expanding universe

The cosmological principle is the foundation for the description of an expanding universe

Once it is accepted that the universe need not be a static entity, but is expanding, it becomes necessary to construct models that are consistent with that new reality. What difficulties might we encounter in undertaking so ambitious a task? First of all, our knowledge of physics has been obtained from experiments carried out on one tiny planet in one infinitesimal region of the universe, in relatively small (human-sized) laboratories. Do the laws we derive hold for the universe at large? We cannot know for certain, but if we accept the cosmological principle they do, although it is possible that they might change with time. A more subtle question is whether there may be laws governing the large-scale behavior of the universe that simply have too small an effect to be seen at the scale of our Earth-bound laboratories. An example would be Einstein’s cosmological constant Λ , which clearly does not affect gravity even on the scale of our galaxy, but might have a profound influence on the universe as a whole. Ultimately, we can only attempt to observe large-scale phenomena, then fit them into our physics. If they cannot be made to fit, then additional physical laws may be required. But unless we are *forced* to think otherwise, we should assume that our existing physics can explain the universe as a whole. We should seek complicating factors that obscure the action of known laws before we postulate new physical principles.

Another significant problem is that the universe is, by definition, unique; we cannot observe other universes. Ideally, we seek a theory that explains everything that happens in this one universe solely in terms of the universe, a theory that describes everything that happens, and does not allow anything that does not occur. We do not have, and may never find, such a theory. In the meantime, we are free to construct simplified models of the universe, to derive their predictions, and to see how well a given model is supported by the observations we can make. How, then, do we derive a model of the universe from Einstein’s theory of general relativity? It is simple enough to describe what must be done: calculate the total matter-energy content of the universe and find the space-time

geometry consistent with that distribution. In practice, this is far from easy.

Before we delve into the details of cosmological solutions to the Einstein equations, let us begin by investigating and clarifying what we mean by expanding space-time, and how such a concept can be incorporated into a space-time metric. By adopting the cosmological principle, which immediately tells us that the universe is homogeneous and isotropic, we place a considerable constraint on the permitted appearance of the metric. Specifically, the metric coefficients must be the same everywhere; they cannot depend either on spatial location or on direction. Recall the usual flat space-time Minkowski metric from special relativity:

$$\Delta s^2 = (c\Delta t)^2 - (\Delta x^2 + \Delta y^2 + \Delta z^2).$$

Since the Minkowski metric coefficients are constants, this provides a trivial example of a homogeneous and isotropic metric. It is, however, a static metric; it does not change with time. We can generalize it by including an arbitrary scale function and allowing that function to vary with the time coordinate:

$$\Delta s^2 = (c\Delta t)^2 - R^2(t)(\Delta x^2 + \Delta y^2 + \Delta z^2). \quad (10.5)$$

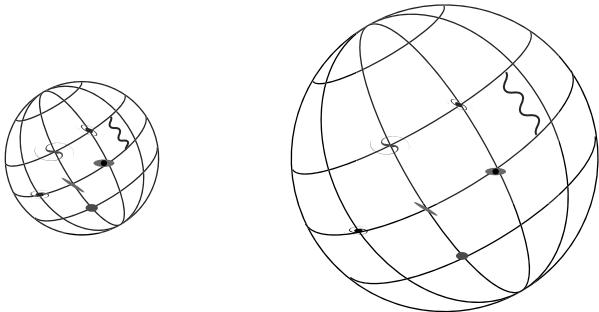
The coefficient $R(t)$ is known as the **scale factor**. The notation means that R is some function that depends only on time t . For example, $R = \text{constant} \times t$ would provide a scale factor that increases linearly with time. We can make space expand (or contract) by adjusting the scale factor function. It is important to understand what we mean here by expansion. If the scale factor increases with time, then two particles, separated spatially at time t_1 by some distance and both at rest with respect to the cosmic frame, are separated at a later time t_2 by a greater distance. The expansion of the universe occurs everywhere and is manifested by an overall increase in the separation of particles; it is not an enlargement into some predetermined, larger entity. Thus it is that a space that may be already infinite can still be expanding.

Let us examine the properties of the cosmological metric of equation (10.5) in some detail. In this form of the metric, the coordinates (x, y, z) are said to be **comoving coordinates**; that is, they remain fixed, and the distance between coordinate locations can be scaled up (expansion) or scaled down (contraction) through the scale function. The surface of a balloon functions well as an analogy. Suppose that we paint lines of latitude and longitude on the surface of the balloon, and then paste some paper disks at various positions on the balloon. Now we inflate the balloon; as it expands, the painted lines expand along with it, as illustrated in Figure 10.9. The coordinates of the paper disks, relative to this grid, do not change. However, the distances between the paper disks, measured along the balloon's surface, do increase. The paper disks themselves do not expand; only the surface of the balloon enlarges. Similarly, the spatial coordinates in the cosmological metric do not change with time; the time variation has been taken entirely into the scale factor $R(t)$. In comoving coordinates, a cluster of galaxies, like the paper

The scale factor $R(t)$ is a measure of relative distance in an expanding universe

The surface of a balloon as an analogy for expanding space

Fig. 10.9 Schematic illustration of a two-dimensional expanding spherical universe, analogous to the surface of a balloon. There is no center of expansion; all points move away from one another. The undulating line indicates the redshifting of light by the expansion of space-time. Comoving coordinates are given by the latitude and longitude lines; these scale up with the expanding sphere.



disks affixed to the balloon, would always keep the same location as time evolves, although the physical distance between one cluster and another would change according to the scale factor. Comoving coordinates are a conceptual aid to help us think about the expanding universe; for real measurements, astronomers would use some coordinate system fixed to a convenient location, such as the center of the Earth or of the Milky Way, and these physical coordinates would change in time. The *metric* measures physical distances; but the *coordinates* that mark locations are arbitrary, and may be chosen for our convenience.

The metric scale factor $R(t)$ then provides all the information we need in order to describe how the universe changes with time. A given model of the universe can be illustrated by a plot of the scale factor versus cosmic time; some examples are shown in Figure 10.10. The vertical axis represents the scale factor R in arbitrary units. In practice, specific values of R are never given. All that matters is how R changes, that is, the ratio of R “now” to R “then.” Figure 10.10(a) shows a scale factor that increases with time, but at an ever-slowing rate. Figure 10.10(b) illustrates a scale factor that increases as a simple linear function of time. Figure 10.10(c) shows a scale factor increasing with time, but at an accelerating rate. The final frame, 10.10(d), displays a scale factor *decreasing* with time. Such a universe is contracting rather than expanding. Graphs such as these are invaluable for clarifying the characteristics of a particular model of the universe.

Expansion is manifest in the large-scale gravity of the universe

Why is it that the scale factor can have an effect on anything, if everything in the universe simply scales up uniformly? If the universe is expanding, then does each galaxy expand as well, and the solar system, and atoms, so that the expansion would be unobservable? There are several answers to this question. First, we must emphasize that in cosmology we consider the largest scales. The standard cosmological solutions to the Einstein equations are obtained by assuming some *averaged* matter or energy distribution. The solutions then correspond to the behavior of the overall gravitational field of the universe. A galaxy, a star, the Earth, an individual human, all reside in local gravitational fields created by the presence of concentrations of mass in a given vicinity; these details are not taken into account in a cosmological solution. Second, in general relativity the metric determines geodesics, and geodesics are the

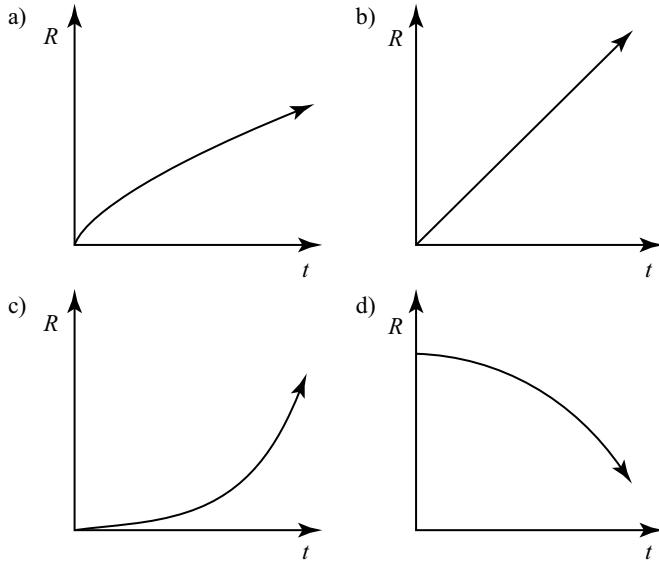


Fig. 10.10 Graphs of scale factor R versus cosmic time t for four representative models. Figures (a), (b), and (c) show expanding universe models, with the rate of expansion slowing, constant, and increasing with time, respectively. Figure (d) illustrates a model that is contracting at an ever-increasing rate.

freefalling worldlines. The geometry of space-time near the Earth, or another massive object, can be regarded as resulting in a gravitational force; similarly, the geometry of the universe provides an expansion tidal force. Objects following the cosmological geodesics move apart. But this expansion tidal force is incredibly weak on the scale of the solar system, or even the galaxy. It only builds up to significance over millions of parsecs. It is very easy for galaxies, or the Earth, or the solar system, to hold themselves together in the face of the expanding universe.

The metric (10.5), which we have introduced to describe the dynamic universe, is really only appropriate for flat space-time geometry. We must introduce a means for the other types of homogeneous and isotropic geometries to be included. Howard P. Robertson and Arthur G. Walker independently constructed such a metric in 1936; they showed that the most general space-time metric for a dynamic, homogeneous, and isotropic universe can be written in the form

$$\Delta s^2 = (c\Delta t)^2 - R^2(t) \left(\frac{\Delta r^2}{1 - kr^2} + \Delta\theta^2 + \sin^2\theta\Delta\phi^2 \right), \quad (10.6)$$

where we use comoving spherical coordinates, with r a radial distance. (This is similar to the use in Chapter 9 of spherical coordinates in describing the Schwarzschild metric.) As was the case for the metric given by equation (10.5), the scale factor $R(t)$ is a function of cosmic time only, but it can change with time in a yet-unspecified manner. This metric is known as the **Robertson–Walker metric**, and its most prominent feature that we have not already encountered is the constant k . This **curvature constant** specifies the curvature of the three-dimensional *spatial* part of the space-time. In our earlier discussion of geometry we stated that there are three homogeneous, isotropic geometries: flat,

The Robertson–Walker metric describes all possible isotropic and homogeneous cosmologies

spherical, and hyperbolic. These correspond to three values of k : $k = 0$ describes a flat geometry; $k > 0$ gives a spherical geometry; and $k < 0$ represents a hyperbolic geometry. We have not yet asserted that our space corresponds to any particular one of these three, only that it is one of them. The Robertson–Walker metric can accommodate all three types of homogeneous and isotropic geometries.

A remarkable feature of this metric is that its geometry is *uniquely* determined by the sign of the constant k . No matter how we adjust our coordinates, the sign of k can never change, although we are able to change its numerical value by a coordinate transformation. It is convenient, therefore, to label the three types of geometry with specific values of k : $k = +1$ corresponds to spherical geometry, $k = 0$ to flat, and $k = -1$ to hyperbolic. It can be shown that the scale factor R is fundamentally related to the magnitude of the curvature of the three-dimensional, spatial part of the metric. If $k = 0$, the spatial geometry is flat and this length is irrelevant to the geometry *per se*, although its change with time controls the behavior of many physical quantities. For the spherical and hyperbolic universes, R indicates the characteristic curvature of space. In the case of spherical spatial geometry, $k = 1$, this scale is easily interpreted as the radius of the spatial part of the space-time at any fixed cosmic time t . For the hyperbolic geometry, which has negative constant curvature, visualization is not possible, but R still sets the length-scale.

Cosmic time corresponds to a clock at rest with respect to the average distribution of matter in the universe

Now let us verify that the Robertson–Walker metric is indeed consistent with the cosmological principle. Is a universe described by such a metric homogeneous? The curvature constant k is the same everywhere; the universe has the same geometry at all points. Every point has the same expansion factor; that is, space-time is evolving in exactly the same way at all spatial locations. We have implicitly included this characteristic by requiring that the scale factor depend only on time, and not on the spatial coordinates. Therefore, this metric describes a homogeneous space-time. Homogeneity makes it possible to define a standard clock for the universe, which can be said to keep **cosmic time**. In the form in which we have presented this metric, the time coordinate t is a cosmic time coordinate. Any clock, anywhere in the universe, that is always instantaneously at rest with respect to the average matter distribution of the universe will keep cosmic time. If such a universe starts with a big bang, then cosmic time indicates the time elapsed since the big bang.

Is this universe also isotropic? The expansion is the same not only everywhere in space, but in all directions, so the metric is isotropic. An isotropic expansion is sometimes said to be *shape-preserving* because shapes do not change, but merely scale up: a square becomes a larger square, a sphere becomes a larger sphere, etc., as the expansion progresses. If the expansion were anisotropic, that is, if the expansion factor were different in one or more spatial directions, then a sphere would be converted into an ellipsoid, and in principle such a transformation is detectable. An anisotropic expansion would also be observable in the redshift pattern; the redshift–distance plots of galaxies in different

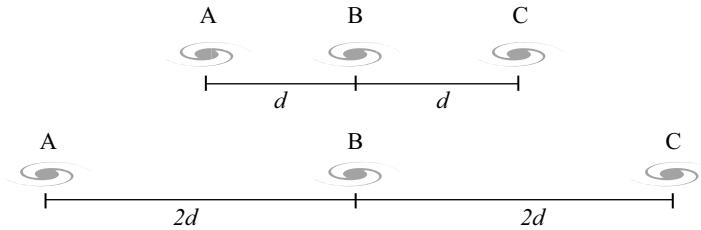


Fig. 10.11 The Hubble law demonstrated by three galaxies, A, B, and C, initially separated by equal distances d . After a time Δt this distance has doubled to $2d$, and the distance between A and C has increased to $4d$. The recession velocity of B from A is $d/\Delta t$, and of C from A is $2d/\Delta t$. The Hubble law predicts just such an increase in the recession velocity with distance.

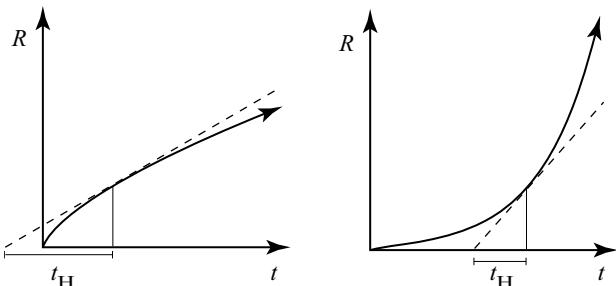
directions would obey different Hubble laws. As best we can tell from current data, the Hubble constant is the same in all directions, so the expansion of the universe appears to be isotropic.

If the universe is isotropic and homogeneous, then it follows that the Hubble law will describe any expansion or contraction. The Hubble law depends only on distance, not on direction; hence it is isotropic. The Hubble law is a simple linear relationship between distance and velocity; a straight line looks the same at all of its points. In fact, the Hubble law is the *only* law for which the expansion is the same in all directions and in all places. We can illustrate the Hubble law with a simple example. Start with a uniform distribution of galaxies, each of which is initially separated from its nearest neighbors by a distance d . Label the galaxies along any one line as A, B, and C, as indicated by Figure 10.11. Now scale everything up by doubling the distances during some time interval Δt . After the expansion, A is separated from B by $2d$ and from C by $4d$. Construct a velocity for each galaxy, relative to A, by computing the distance moved and dividing by the time interval. Thus B moves away from A at a recession velocity of $(2d - d)/\Delta t = d/\Delta t$, and C recedes at $(4d - 2d)/\Delta t = 2d/\Delta t$. The recession velocity of any galaxy relative to A increases linearly with the distance from A, which is the Hubble law. However, our choices of reference galaxy and line of sight were completely arbitrary; we could have centered our reference on any of the galaxies and looked in any direction. From this we see that a homogeneous and isotropic expansion is described by the Hubble law.

The Hubble law is a statement of how the universe expands, and the Hubble constant is a measure of how fast it expands. In Chapter 11 we will learn how the Hubble constant can be related to the change with time of the scale factor. For now, we shall accept that the Hubble constant indicates the speed of expansion. An extremely useful byproduct of this fact is that the Hubble constant can tell us something about the length of time over which the galaxies have been separating. From the definition of the Hubble law, equation (10.4), it is clear that the unit of the Hubble constant is inverse time (s^{-1}). Therefore, the inverse of the Hubble constant is an interval of time. This interval is called the **Hubble time** (or Hubble period, or Hubble age). Mathematically, the

The Hubble law is appropriate for an isotropic and homogeneous universe

Fig. 10.12 Relation of the Hubble time, which is a linear extrapolation from time “Now” back to $R = 0$, as indicated by the straight dashed lines, to the actual age of the universe, obtained from the scale factor. On the left, a universe for which the rate of expansion is decelerating; in this case, the Hubble time is longer than the true age of the universe. On the right, a universe whose expansion is accelerating; its Hubble time is shorter than its actual age.



Hubble time is simply

$$t_H = \frac{1}{H}, \quad (10.7)$$

The Hubble time is an estimate of the age of the universe

where H is expressed in consistent units, such as inverse seconds. This equation shows that the higher the expansion speed, that is, the larger the Hubble constant, the shorter the Hubble time. This should make intuitive sense, as well. If the universe is expanding rapidly (large Hubble constant), then it would reach its present scale more quickly than it would at a small expansion rate (small Hubble constant). For reference, if the Hubble constant is approximately $100 \text{ km s}^{-1} \text{ Mpc}^{-1}$, then the Hubble time is roughly 10^{10} years; if $H = 50 \text{ km s}^{-1} \text{ Mpc}^{-1}$, then $t_H = 2 \times 10^{10}$ years. Because the relationship between the Hubble constant and the Hubble age is a simple inverse proportionality, Hubble times for other values of the Hubble constant can be easily computed by using these reference values.

The Hubble time provides an *estimate* of the age of an expanding universe. More exactly, it is the age of an idealized universe that expands from zero size at a constant rate given by the value of H in question. We shall soon discover that the Hubble constant is not, in general, truly a constant, but varies over the history of the universe; thus the Hubble time gives not the actual age of the universe, but an approximate age. The Hubble time is akin to an estimate of the time required for a journey. If you were to take a car trip, at any point in your trip you could use the instantaneous speed on the speedometer, plus the distance traveled as measured by the odometer, to estimate how long you had been driving. If your speed had been changing, then obviously this could be a rather poor estimate, whereas if the speed were relatively constant, it would be a reasonably good estimate. On your car ride, if you wanted to compute from your speedometer readings the exact time you had been driving, you would have to know in detail how your speed varied during your entire trip. Similarly, to find the exact age of the universe, we must know how the Hubble constant has changed with time. Given a model, we can compute the precise age of this universe. In the absence of sufficient data to decide which model best describes the universe, the Hubble age provides a useful and convenient timescale.

A related concept is the **Hubble length**. This is the distance that light could travel in one Hubble time,

$$D_H = ct_H = \frac{c}{H}. \quad (10.8)$$

The Hubble length is an estimate of the size of the observable universe

For a Hubble constant of $50 \text{ km s}^{-1} \text{ Mpc}^{-1}$, this length is 20 billion lightyears. The Hubble length has several interpretations. Perhaps the easiest way to understand its significance is to consider a sphere around the Milky Way of radius equal to the Hubble length; this imaginary delimiter is called the **Hubble sphere**. The volume enclosed by the Hubble sphere is an estimate of the volume of the universe that could possibly be within our past lightcone. The Hubble sphere thus is not the entire cosmos; it estimates the limit of our *observable* universe, containing all that could have affected us, or been affected by us, up to the present time. *Every* point in the universe has its own Hubble sphere; the existence of the Hubble sphere thus does not imply a center at any special location. The Hubble length can change with cosmic time if the Hubble constant changes; the exact time behavior depends upon the variations in the Hubble age at different epochs. If the expansion rate is slowing, then the Hubble age increases with cosmic time; hence the Hubble sphere will become larger with time.

Another interpretation of the Hubble length is that it is that distance at which the speed of recession from our vantage point is equal to the speed of light. Again this shows that the Hubble length represents the greatest distance to which we can possibly see, regardless of any other limitations under which might operate. Just as the Hubble time is only an approximation to the age of the universe, with the actual age dependent upon the exact model, so the Hubble sphere is only an approximate measure of the size of the observable universe. The true size must be computed using a specific model $R(t)$. The Hubble sphere is still a good approximate length-scale for conceptualizing the universe, and serves as a perfectly adequate estimate for most purposes.

The cosmological redshift

Now that we have developed the Robertson–Walker metric, we can understand the origin of the cosmological redshift. Because of the change in the cosmic scale factor, the wavelength at the time the light was emitted is different from the wavelength at the time it is detected. The cosmological redshift is a consequence of the effects of gravity upon space-time, as specified by a general-relativistic metric. (As with any metric, two events in space-time connected by a light beam have $\Delta s = 0$.) As lengths increase with time in the expanding universe, the wavelengths of light moving through space also increase. Specifically, since wavelength λ is just a length, and we know from the scale factor $R(t)$ how lengths change in a Robertson–Walker metric, we can write for the wavelength “now,”

λ_{rec} , the formula

$$\lambda_{\text{rec}} = \left(\frac{R_{\text{now}}}{R_{\text{then}}} \right) \lambda_{\text{em}}. \quad (10.9)$$

The cosmological redshift is directly related to the scale factor $R(t)$

But the redshift, from whatever cause, is *defined* as the change in wavelength divided by the standard value of that wavelength, that is, $z = (\lambda_{\text{rec}} - \lambda_{\text{em}})/\lambda_{\text{em}}$, where λ_{em} is the wavelength as measured at the emitter. Thus we find

$$1 + z = \frac{\lambda_{\text{rec}}}{\lambda_{\text{em}}} = \frac{R_{\text{now}}}{R_{\text{then}}}. \quad (10.10)$$

The redshift provides a direct measure of the ratio of the scale factor “now” to the scale factor “then,” when the light was emitted.

Relativity has shown that anything causing a change in lengths (such as length contraction) must also affect time as well (time dilation). Wavelength times frequency always equals the speed of the wave; light traveling through a vacuum obeys the equation $\lambda\nu = c$. Thus both the wavelength and the frequency ν change with the expansion, since c is a constant for all observers and all times. Frequency is measured in cycles per second, so the inverse of a frequency is a time. Hence the expanding universe produces a cosmological time dilation, as well as a redshift. For example, suppose that a clock attached to a quasar with a very large redshift measures the frequency of a particular ultraviolet light beam, emitted somewhere within the quasar, as 10^{15} cycles per second. By the time that light reaches us on Earth, it has shifted into the infrared, with a frequency of 10^{14} cycles per second as we measure it. Thus it takes 10 of our seconds for us to see the quasar clock tick off one quasar second. In other words, we see the quasar clock running very slow. We can derive this using the wavelength–frequency relationship, $1/\nu = \lambda/c$, in combination with the redshift formula, equation (10.10), to obtain

$$\frac{\nu_{\text{em}}}{\nu_{\text{rec}}} = \frac{\lambda_{\text{rec}}}{\lambda_{\text{em}}} = \frac{R_{\text{now}}}{R_{\text{then}}} = 1 + z. \quad (10.11)$$

Notice that whereas an expanding universe *increases* wavelengths, it *decreases* frequencies. The frequency formula is, as expected, very similar to the formula for wavelength; it can be used to indicate the time dilation of very distant sources relative to our clocks. In the above example, the quasar has a redshift $z = 9$; a more typical quasar might have a redshift of $z = 1$. An observer on Earth would find that this quasar’s clocks run slow, relative to Earth clocks, by a factor of two.

We have seen how two important observable properties of the universe are related to the scale factor: the cosmological redshift and the Hubble law. These effects contain subtleties that lend themselves to misunderstandings. One point of confusion is the belief that the cosmological redshifts are ordinary Doppler shifts resulting from motions of the galaxies through space. After all, cosmologists talk about “expansion velocities,” and the Hubble law relates just such a velocity to a distance. An analogy with Doppler shifts can be useful, particularly for nearby galaxies where the redshifts are small, but the cosmological

Cosmological time dilation

redshift is really more akin to the gravitational redshift discussed in connection with black holes. Both cosmological and gravitational redshifts arise directly from a metric. The cosmological redshift is produced by photons traversing expanding space. It does not occur because of relative motion at the moment of the emission or reception of the light, as is the case for the conventional Doppler shift, but as the light *travels* through space. The formula $z + 1 = R_{\text{now}}/R_{\text{then}}$ tells us only the *ratio* of the scale factor today, when the light is received, compared to the scale factor at the time the light was emitted. It tells us nothing about how the expansion (or contraction) proceeded with time.

As an illustration, suppose that the universe is not expanding at some arbitrary time t_{then} . A distant galaxy emits light toward the Milky Way at t_{then} , while it is at rest with respect to the Milky Way. Write the scale factor at that time as R_{then} . Suppose further that after the light was emitted, the universe begins to expand rapidly. As the light crosses the space between the emitting galaxy and the Milky Way, it will be redshifted because of the expanding space. Finally, suppose that the expansion stops abruptly at a scale factor R_{now} , just as the light reaches us. When the light arrives it still has a large redshift, in accordance with equation (10.10), despite its being both emitted and received while the source is at rest with respect to the Earth. It would have had exactly the same redshift in the more realistic case that the universe had expanded continuously from R_{then} to R_{now} . Again, the cosmological shift observed from some specific galaxy tells us only the relative scale factors, not the way in which the universe evolved from R_{then} to R_{now} .

This is not to say that we cannot derive such information about our universe, but it is a difficult business. Redshift alone is not enough; we need another piece of information. Consider equation (10.4), the Hubble law. We called it the theoretical Hubble law because it describes the state of the universe at one instant of cosmic time, and this is something that we *cannot* directly observe. When we make observations, we do not see the universe at a single moment in cosmic time. Because the speed of light is finite, when we look at distant stars and galaxies we are seeing them as they were at the time in the past when that light left them. The farther we look into space, the farther back in time we look; the travel time of the light is called the **lookback time**. We see a redshift because the universe had a different scale factor when the light left the emitter. The redshift gives us the ratio $R_{\text{now}}/R_{\text{then}}$, and the lookback time would tell us how far in the past the universe had the scale factor R_{then} . If we knew both the lookback time and the redshift for a large number of objects, we could construct a complete graph of $R(t)$. In practice, the redshift is easy to measure, but the lookback time is not. To obtain the lookback time, we must measure the distance to the object emitting the light. This brings us right back to the same difficult question that confronted Hubble, namely, the impediments to seeing distinct standard candles at such great distances. When we compute the distances to remote objects, we must take into account the change of the Hubble constant with time, and the increasing distances

Cosmological redshifts are not Doppler shifts

Both independent distances and redshifts are required to determine $R(t)$

between objects as time progresses. Only when we look at nearby galaxies can we ignore the complexities of a specific model. For such galaxies, the lookback time is relatively small, and the Hubble constant has not changed significantly since their light was emitted. In such cases, we can approximate distances directly by relating the Hubble law to the redshift. But this is only an approximation, and a good approximation only for nearby galaxies. For anything else, we need a complete model $R(t)$.

We have learned how the expanding universe was discovered observationally. We have seen how the expanding universe can be incorporated within a space-time metric through the use of a homogeneous and isotropic scale factor. Finally, we have examined how some of the observed properties of the universe, such as the Hubble constant and the redshifts, can be related to the all-important scale factor. There remains the task of constructing specific models of $R(t)$; we take up this challenge in the next chapter.

Chapter Summary

This chapter recounts several important historical threads in the development of modern cosmology. Controversy over the nature of spiral nebulae had persisted since the late 18th century, with one camp insisting they were external “universes,” while their opponents were equally convinced that the spiral nebulae were localized clusters of stars within our Galaxy. An important early discovery was Shapley’s determination of the size of the Milky Way Galaxy, and of our location within it. Shapley found that the Milky Way is much larger than previously believed, and on this basis he erroneously concluded that the spiral nebulae must be relatively nearby clusters. Shapley and Curtis participated in a famous debate in 1920 over the nature of the spiral nebulae, but insufficient data prevented a resolution of the puzzle. Finally, Edwin Hubble determined that the Andromeda Nebula (now known as the Andromeda Galaxy) was much too distant to lie within the confines of the Milky Way; Hubble had discovered external galaxies. In the first quarter of the 20th century, humanity’s view of the cosmos leaped from a fairly limited realm of the Sun surrounded by an amorphous grouping of stars, to one in which the Milky Way is just a typical spiral galaxy in a vast universe filled with galaxies.

Not long after Hubble’s discovery of external galaxies came his discovery of a linear relationship between their redshifts and their distances, a relationship known today as the Hubble law. The value of the constant of propor-

tionality, the Hubble constant, is one of the fundamental cosmological parameters. The Hubble “constant” is not really constant because in general it changes with time, although at any given instant of cosmic time in a homogeneous, isotropic universe, it is the same at all spatial locations. The inverse of the Hubble constant, called the Hubble time, gives an estimate of the age of the universe. The distance that light could travel in a Hubble time is called the Hubble length. A sphere can be defined with the Hubble length as its radius; this sphere is called the Hubble sphere, and it gives an estimate of the volume of the universe that is observable by us.

Measuring Hubble’s constant requires accurate distances to increasingly remote galaxies. This can be done by measuring the apparent brightness of a standard candle, that is, an object of known luminosity. One of the best such standard candles is the Cepheid variable star. The *HST* has now been able to detect Cepheid variable stars in the Virgo galaxy cluster. Several Cepheids have been observed, and this new data give us a Hubble constant of about $72 \text{ km s}^{-1} \text{ Mpc}^{-1}$. Cepheids in turn have been used to calibrate an even brighter standard candle, the Type Ia supernova.

The development of the theory of general relativity provided the framework in which Hubble’s discovery could be understood. Einstein found that his equations would not admit a static, stable model of the universe, even with the addition of the cosmological constant. The timely

discovery of the redshift–distance relationship provided evidence that the universe was *not* static, but was expanding. The Robertson–Walker metric is the most general metric for an isotropic, homogeneous universe that is also dynamic, that it, it changes with time. An im-

portant parameter in this metric is the scale factor, $R(t)$, the quantity which describes how lengths in the universe change with cosmic time. The scale factor is directly related to the cosmological redshift, the change in wavelength of light as it traverses the universe.

Key Term Definitions

redshift A shift in the frequency of a photon toward lower energy.

blueshift A shift in the frequency of a photon toward higher energy.

cosmological redshift A redshift caused by the expansion of space.

nebula A cloud of gas in space.

nova An abrupt, very bright flare-up of a star. Most likely due to the accumulation of hydrogen from a companion upon the surface of a white dwarf. The pressure and temperature grow in this accumulated matter until a thermonuclear explosion is generated.

supernova The explosive death of a star. Type Ia supernovae occur when a white dwarf accumulates too much gas from a companion upon its surface, causing it to exceed the Chandrasekhar limit. Type II supernovae occur when a massive star has reached the endpoint of nuclear fusion and can no longer support itself. In both cases, the result is a catastrophic gravitational collapse, and an explosion so violent that elements heavier than iron are created. Any remaining core becomes a neutron star or a black hole.

cosmic distance ladder The methods by which increasing distance is measured in the cosmos. Each depends on a more secure technique (or “rung”) used for smaller distances.

parallax The apparent shift in the position of a celestial object, such as a star, due to the changing vantage point of the observer. Astronomical parallax can be caused, for example, by the orbital motion of the Earth, or its daily rotation (*durnal parallax*).

parsec That distance producing one second of arc of parallax over the baseline of the Earth’s orbit. One parsec (pc) corresponds to 3.26 lightyears.

luminosity distance The inferred distance to an astronomical object derived by comparing its observed brightness to its presumed total luminosity.

standard candle An object of known intrinsic luminosity, useful in the measurement of luminosity distances.

extinction In astronomy, the removal of light from a beam by whatever means, such as absorption and scattering.

Cepheid variable A type of variable star whose period of variation is tightly related to its intrinsic luminosity.

Hubble constant The constant of proportionality (H) between recession velocity and distance in the Hubble law. It is not actually a constant, because it can change with time over the history of the universe.

Hubble law The relationship between recession velocity and distance, $v = H\ell$, for an isotropic, expanding universe.

Tully–Fisher relation An empirical relationship between the width of the 21-cm line of hydrogen emissions from spiral galaxies and the mass of the galaxy. The relationship arises because a larger mass increases the rotation rate, and a faster rotation causes a broader line; the precise calibration must be determined observationally.

Hubble expansion The separation of galaxies due only to the expansion of space.

peculiar velocity The unique velocity of an object such as a galaxy, due to its individual gravitational interactions with other objects, not due to the general cosmological recession.

cosmological constant A constant introduced into Einstein’s field equations of general relativity in order to provide a supplement to gravity. If positive (repulsive), it counteracts gravity, while if negative

(attractive), it augments gravity. It can be interpreted physically as an energy density associated with space itself.

dark energy Energy associated with the vacuum of space that drives acceleration in the overall expansion of the universe.

de Sitter model A model of the universe that contains no matter, but only a positive cosmological constant. It expands exponentially forever.

scale factor The quantity (R) that describes how the scale changes in the expanding (or contracting) universe.

comoving coordinates Coordinates fixed with respect to the overall Hubble flow of the universe, so that they do not change as the universe expands.

Robertson–Walker metric The metric that describes an isotropic and homogeneous cosmological space-time.

curvature constant A constant (k) appearing in the Robertson–Walker metric which determines the curvature of the spatial geometry of the universe.

cosmic time A time coordinate that can be defined for all frames in a homogeneous metric, representing the proper time of observers at rest with respect to the Hubble flow. In a big bang model this coordinate marks the time elapsed since the initial singularity.

Hubble time The inverse of the Hubble constant, $t_H = 1/H$. The Hubble time, also called the Hubble age or the Hubble period, provides an estimate for the age of the universe.

Hubble length The distance traveled by light along a straight geodesic in one Hubble time, $D_H = ct_H$.

Hubble sphere A sphere, centered about any arbitrary point, whose radius is the Hubble length. The center of the Hubble sphere is not a “center” to the universe, because each point has its own Hubble sphere. The Hubble sphere approximately defines that portion of the universe that is observable from the specified point at a specified time.

lookback time The time required for light to travel from an emitting object to the receiver.

Review Questions

- (10.1) In retrospect, it seems obvious that the spiral nebulae are external galaxies. Discuss the reasons that this hypothesis was so slow in gaining acceptance. What finally proved it?
- (10.2) You are measuring distances to galaxies using a particular standard candle. At a professional meeting, another astronomer announces that your standard candle is actually twice as luminous as previously believed. If she is correct, how would you have to modify your derived distances?
- (10.3) Discuss the difficulties in measuring extragalactic distances. What phenomenon confused scientists for several decades, causing them to overestimate distances? What kinds of distance indicators can be used? Describe some potential sources of error with the methods.
- (10.4) Assume that the Hubble constant is $50 \text{ km s}^{-1} \text{ Mpc}^{-1}$. Using the Hubble law and the nonrelativistic redshift formula ($z = v/c$, where z is the redshift), calculate the distance of a galaxy whose spectrum has a redshift of one percent. (The speed of light is $3 \times 10^5 \text{ km s}^{-1}$.)
- (10.5) Explain why the overall expansion of the universe does not make the solar system expand as well.
- (10.6) A Hubble constant of $50 \text{ km s}^{-1} \text{ Mpc}^{-1}$ corresponds to a Hubble time of about 20 billion years. What would be the Hubble time for a Hubble constant of $100 \text{ km s}^{-1} \text{ Mpc}^{-1}$? $75 \text{ km s}^{-1} \text{ Mpc}^{-1}$? If the rate of expansion of the universe is slowing down with time, will the Hubble time over- or underestimate the age of the universe?
- (10.7) The Hubble sphere forms a sort of edge to our observable universe. Why isn't this a real edge to the universe? Why doesn't this edge violate the cosmological principle or the Copernican principle?
- (10.8) In order to prevent his model from collapsing, Einstein added a term Λ to his model. A positive Λ resulted in a repulsive force to counteract gravity. What do you think would happen if a negative Λ were used instead?

- (10.9) Some quasars have redshifts of 4 or even greater. Redshifts can be caused by relative motions, by gravitational fields, or by the expansion of the universe. Consider the case of gravitational redshift. Using equation (9.7), compute how close to the black hole the emitting region would have to be in order to produce a redshift $z = 4$. Give your result in terms of the fraction of the Schwarzschild radius. Why is it unlikely that the redshifts of quasars could be explained solely by gravitational redshifts? (10.11)
- (10.10) Consider a quasar at a redshift $z = 2$. If the quasar's light output varies with a period of one day as we observe it, what is the period of varia-

tion in the quasar's frame? What does the quasar's *lookback time* mean?

Consider the universe at a redshift of $z = 2$. If two galaxies were separated by a distance l at that time, what is their separation today? What is their separation in *comoving coordinates* today compared with then?

Draw a graph showing scale factor versus time using the following data for redshifts and lookback times: At $z = 0$, $t_{lb} = 0$; $z = 1/3$, $t_{lb} = 12$; $z = 1$, $t_{lb} = 21$; $z = 1.5$, $t_{lb} = 24$; $z = 3$, $t_{lb} = 28$; $z = 9$, $t_{lb} = 31$. The units are arbitrary. Locate these data points on the graph and then draw a freehand curve through them.

This page intentionally left blank

Modeling the Universe

11

The most incomprehensible thing about the world is that it is comprehensible.

Albert Einstein

Two major developments, Hubble's observations and Einstein's theory of general relativity, moved the subject of cosmology out of the realm of the mainly philosophical and firmly into the arena of science. Now the task of developing an actual model of a dynamic universe lies before us. While at first glance it may seem an impossible undertaking, we are aided by the adoption of the cosmological principle. This is an enormous simplification, for it implies that the metric which describes space and time, and specifies their evolution, must be the Robertson–Walker metric. The real universe is complex, with many intricate structures and objects, so it may seem excessively crude to assume that the universe is perfectly smooth and homogeneous; but we must start somewhere, and it seems prudent to begin with the simplest case. As yet, however, we have gone no further than this, and until we determine the parameters $R(t)$ and k , we can say nothing more. How might we go about evaluating the scale factor and the curvature constant?

The construction of cosmological models corresponds to the task of solving Einstein's equations of general relativity for an isotropic and homogeneous universe. Einstein's equations (8.5) state that the geometry of the universe is determined by its mass and energy content. The detailed mass and energy distribution of the universe is obviously very nonuniform. Galaxies and stars are scattered unevenly throughout space; interstellar and intergalactic gas exist in irregular patterns. In addition to the matter distribution, the universe is filled with photons as well as possibly other, more exotic, particles, all contributing some quantity of energy. In keeping with the cosmological principle, however, we assume that the characteristics of individual clumps of matter, such as the galaxies, are not important. Instead, we shall take all the galaxies, stars, and planets in the universe, grind them up into a fine dust, and distribute that dust evenly throughout space, so that at every location the average mass density takes a constant value ρ . The universe may also be filled with energy from sources other than rest mass-energy; similar to the mass-energy, these other forms can be characterized by a uniform energy per unit volume, or *energy density*. By replacing the actual matter and energy constituents of the universe with these con-

Key Terms:

- Friedmann equation
- standard model
- big bang
- big crunch
- matter density parameter
- critical density
- deceleration parameter
- closed universe
- flat universe
- Einstein–de Sitter model
- open universe
- cosmological constant
- lambda density parameter
- de Sitter model
- steady state model
- Lemaître model

A model of the universe is a description of $R(t)$.

stant average values, we greatly simplify the right-hand side of Einstein's equations. The variables describing the contents of the universe, such as ρ , are now independent of spatial location; at most they can depend only upon time, for consistency with the cosmological principle.

We shall go no further here with Einstein's equations; solving for the metric term from the geometry term requires more differential calculus than we wish to demand. Fortunately, we can gain considerable insight into the nature of the solution by considering Newtonian physics in a flat space. In fact, Newtonian physics is an adequate description of the universe as long as the distances we consider are much less than the radius of the Hubble sphere, and the expansion velocities are much less than c . In a sufficiently small region, we can safely ignore any curvature of space, so the assumption of a flat space is not a severe limitation. Moreover, in a homogeneous, isotropic universe, anything we learn locally tells us something about the way the universe works everywhere. Of course, although these may seem to be reasonable justifications for using Newtonian physics, cosmologists do find their solutions for the universe using Einstein's general relativity. Remarkably, the equations that result from the Newtonian analysis are almost exactly the same as those that emerge from the Einstein equations.

A Newtonian universe

The universe envisioned by Newton was infinite and unchanging, filled in all directions with stars acting under the mutual gravitational forces of all the other stars. The image is much like that of the air molecules in a huge room; the molecules fly around with random velocities and interact with one other, but overall the air in the room is still. There was one aspect of this model that was rather troubling. If the cosmos is filled with stars that attract one other by gravitational forces, should not all the stars pull themselves together into a single clump? Newton reasoned that in an infinite universe, the forces would be balanced at every point due to an equal gravitational attraction in each direction. But in a universe with an infinite number of stars, the forces in all directions would be infinite, so this would be a delicate balance indeed!

Rather than dealing with an infinite extent of stars, let us confine our attention to a large, but finite, spherical portion of the universe, with radius R . Assume that it is filled uniformly with matter. We shall focus our attention on a small bit of matter sitting on the edge of the sphere. Let this little "test particle" have a mass m_t , while the total mass of all the other matter within the sphere shall be M_s , as illustrated in Figure 11.1. There is nothing special about this test particle; indeed, since R is arbitrary, as long as it meets the requirements for a Newtonian analysis to be valid, the test particle could actually represent *any* random gravitating particle within the distribution of matter. So long as any exterior matter is distributed uniformly, the gravitational effects of the matter *outside* the sphere cancel out, in a manner similar to Newton's

The universe as a sphere of dust.

explanation of the gravitational attraction due to a sphere.¹ Therefore, the only contribution to the gravitational acceleration of our test particle comes from the matter within the sphere. We can now apply Newton's formula for the gravitational force,

$$F_g = -\frac{GM_s m_t}{R^2}, \quad (11.1)$$

to obtain the net gravitational force acting on the test particle. The usual Newtonian gravitational acceleration,

$$g = -\frac{GM_s}{R^2}, \quad (11.2)$$

directed toward the center, acts on the particle. The situation we have described is exactly analogous to the force of gravity on a human body (a test particle) due to the mass of the Earth (a large mass-filled sphere). Just as on the Earth, the acceleration due to gravity is the same for all test particles and behaves as if all the mass of the sphere were concentrated at its center.

Imagine now that the test particle is simply one particle at the edge of a sphere of dust, located at radius R . Assume that the sphere of dust can expand or contract, though its total mass M_s cannot change. Hence the quantity R is not only the radius of the sphere, but also specifies the location of the test particle. Assume now that the particle and sphere are expanding with some outward velocity. A velocity is the change of position with time, so we can write

$$v = \frac{\Delta R}{\Delta t} \equiv \dot{R}. \quad (11.3)$$

We have introduced a new notation, \dot{R} , to represent the change in the location R with respect to time. In calculus the dot notation represents a time derivative, $\dot{R} = dR/dt$. The “double dot” notation, \ddot{R} , is the second time derivative. For exactness, we shall employ this notation, although for the present purposes simply think of \dot{R} as a velocity and \ddot{R} as an acceleration. The velocity too will change with time due to gravitational acceleration. We write

$$\frac{\Delta v}{\Delta t} \equiv \ddot{R} = g = -\frac{GM_s}{R^2}, \quad (11.4)$$

where we have used equation (11.2) for the acceleration. As the sphere expands outward, gravity will reduce its velocity, possibly leading to an infall and collapse.

We would like to solve equation (11.4) to determine the location R as a function of time t . This is a differential equation and is solved using integral calculus. Rather than solving it directly, however, we shall instead describe the solution's physical behavior. Our system consists of

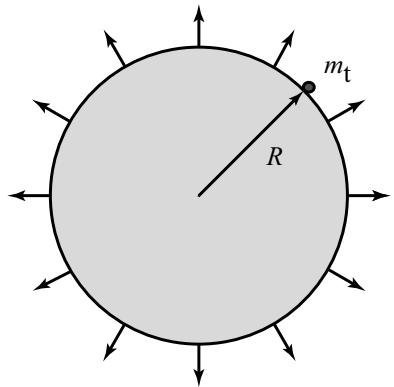


Fig. 11.1 A Newtonian universe. A sphere of radius R contains uniformly distributed dust. A test particle with mass m_t , located at the edge of the sphere, feels the gravitational attraction of all this mass directed toward the center of the sphere. The whole system is expanding due to some initial outward velocity, but gravity causes that expansion to decelerate.

¹Chapter 3 describes Newton's proof that a sphere attracts gravitationally as if all its mass were at its center.

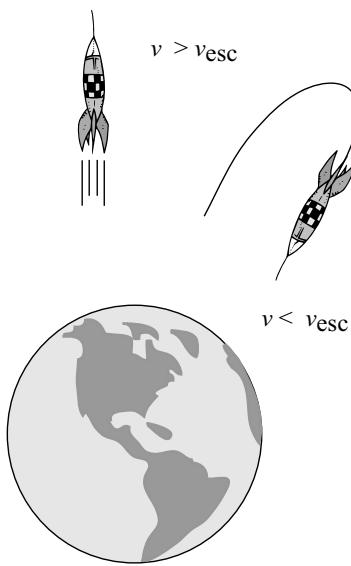
Escape velocity

Fig. 11.2 To leave the gravitational field of the Earth, a rocket's velocity must exceed the escape velocity.

a sphere of particles moving in the radial direction under the influence of gravity; the “test particle” can be regarded as simply any representative particle. We have encountered such a situation before, in another guise. In Chapter 9 we discussed the escape velocity, defined to be that velocity which just permits a particle to escape from the gravitational field of an object of a certain mass and radius,

$$v_{\text{esc}} = \dot{R} = \sqrt{\frac{2GM_s}{R}}. \quad (11.5)$$

The concept of escape velocity has its most obvious application when considering the launch of a rocket from the Earth. If a rocket has less than escape velocity, it cannot escape the Earth but falls back. A rocket launched radially precisely at the escape velocity will be slowed by gravity, such that at any distance R from the Earth, its speed equals the escape velocity appropriate to that R ; the rocket travels indefinitely outward with a velocity that approaches zero as R goes to infinity. Finally, if the rocket's velocity exceeds escape velocity, then it has more than enough speed to leave the Earth's gravitational field. The rocket escapes to infinity, where it still has a positive velocity. The question in the case of the expanding, self-gravitating sphere is whether or not the particles that make it up are moving outward with sufficient velocity to avoid recollapsing due to their own gravity. If the particles have *precisely* the correct velocity just to avoid recollapsing as R goes to infinity, the expansion speed will be equal to the escape velocity at every R .

We can also cast the issue more precisely by phrasing it in terms of the energy of motion, or the *kinetic energy*, of the test particle. In the Newtonian physics we are considering for the moment, this energy is given by one half the mass of the particle times its velocity squared, $E_k = \frac{1}{2}m_t v^2$, and the kinetic energy per unit mass is simply this quantity divided by the mass m_t , that is, $\mathcal{E} = \frac{1}{2}v^2$. The square of the escape speed divided by 2 thus gives an escape energy per unit mass. We can generalize the problem by noting that among all possible expansions, movement precisely at the escape speed is a special case; in general, the speed can be less than or greater than this critical case. Thus we can add an additional constant term, positive for speed exceeding, negative for speed less than, and zero for energy precisely at the escape value. The generalization of equation (11.5) can then be written in the form

$$2\mathcal{E} = \dot{R}^2 = \frac{2GM_s}{R} + \text{constant}. \quad (11.6)$$

We can evaluate the constant by allowing the radius to go to infinity, causing the “matter” term to vanish; we find that this constant is equal to twice the kinetic energy per unit mass remaining when the sphere has expanded to infinite size. We shall denote this quantity by E_∞ . We can thus express equation (11.6) as

$$\dot{R}^2 = \frac{2GM_s}{R} + 2E_\infty. \quad (11.7)$$

The quantity E_∞ as defined here can be zero, positive, or negative.

Given an expanding Newtonian sphere, we consider each of the possibilities in turn:

- (1) If the energy per unit mass at infinity is negative, $E_\infty < 0$, then the sphere has net negative energy. Its expansion will halt at some point *before* it reaches infinite radius. It will then recollapse.
- (2) If the energy per unit mass at infinity is zero, $E_\infty = 0$, then the sphere has zero net energy. It has exactly the velocity required to keep expanding forever, although the velocity will tend to zero as time and radius go to infinity.
- (3) If the energy per unit mass at infinity is positive, $E_\infty > 0$, then the sphere has net positive energy. It will keep expanding forever at a rate that is faster than necessary to prevent it from stopping and recollapsing, and will reach infinite radius with some velocity remaining.

Now we must make the great leap from the analogy of the sphere to the universe. Rather than regarding the sphere as an isolated ball of mass, we can regard it as a typical volume of space, filled with a smooth distribution of dust at a constant density. Next, we will rewrite our equations in terms of density, not total mass, because the average density of the universe is locally measurable, whereas the total mass is not. Density is mass per unit volume; thus we shall replace the total mass within the sphere by its volume multiplied by its density, $M_s = (4/3) \pi R^3 \rho$. Performing this substitution for the acceleration equation (11.4) we obtain

$$\ddot{R} = -\frac{4}{3}\pi G\rho R \quad (11.8)$$

and for the velocity equation (11.7) we derive

$$\dot{R}^2 = \frac{8}{3}\pi G\rho R^2 + 2E_\infty. \quad (11.9)$$

What is the limit of this equation as R becomes large without bound? Although we have now written the equation in terms of density, mass is still conserved, meaning that the quantity ρR^3 remains constant. Equations (11.7) and (11.9) behave in exactly the same way as the scale R becomes infinitely large. Thus, like the Newtonian sphere, the universe too can have positive, zero, or negative net energy. If the universe has negative energy at infinity, it cannot “escape” its own gravity, and it will eventually cease to expand, subsequently collapsing back on itself. If the universe has positive energy at infinity, it has sufficient energy to expand forever. If the universe has zero energy at infinity, it will expand forever, but the expansion velocity will drop to zero as the radius of the universe becomes infinite. The constant $2E_\infty$, then, is related to the fate of the universe.

Three possible fates for the expanding sphere

The standard models

So far our discussion has been within the context of Newton's equation of gravity. We know from our study of general relativity that Newtonian gravity cannot apply to the universe as a whole. We were justified in applying it to investigate a spherical region carved out of the universe, as long as the sphere was relatively small compared to the Hubble length and any velocities within it were nonrelativistic. Under such conditions, Newtonian gravity is a good approximation for gravity in a flat space. Of course, to extend our analysis to the entire universe and to curved spaces, we must return to general relativity. If we had worked out the fully relativistic equations for a universe with uniform matter density, we would have obtained the equation

$$\dot{R}^2 = \frac{8}{3}\pi G\rho R^2 - kc^2 \quad (11.10)$$

as the relativistic equivalent of equation (11.9), with $-kc^2$ filling the role of the energy at infinity. Here k is the same curvature constant that appears in the Robertson–Walker metric. Since we can always choose our coordinate values to take whichever scale we wish, we shall adjust them so that k will take one of the three values 0, +1, or −1. These correspond to the three isotropic and homogeneous geometries: flat, spherical, and hyperbolic, respectively.

Equation (11.10) is nearly the same as that which emerged from our Newtonian treatment of the universe, except that now R is the scale factor rather than the radius of some arbitrary sphere. In other words, the gravity of the mass in the universe acts on the space-time scale factor in much the same way that the gravity of the mass inside a uniform sphere acts on its own radius R . There are important additions due to relativity, however. First, we have replaced the E_∞ term in the Newtonian equation with the curvature constant. This term retains its significance as an energy at infinity, but now it is tied into the overall geometry of space. Second, relativity requires that we must include *all* forms of energy, not just rest mass, in our definition of ρ ; mass and energy are equivalent, and both contribute to the force of gravity. This relativistic equation (11.10) is called the **Friedmann equation**, after its developer Alexander Friedmann.

Because of the complicating addition of energy to the source of the gravitational field, we cannot solve the Friedmann equation until we have specified how the total mass-energy density changes with time. This requires two more equations: the relativistic equation for the *conservation* of mass-energy density, and an *equation of state*, a relationship between matter density and energy density. We shall not derive these equations; instead, we shall content ourselves with describing some important qualitative features. In the case of ordinary matter density, the quantity ρR^3 remains constant; as R increases, ρ decreases precisely as in the previous Newtonian analysis. In our present universe, the only significant contribution to nonmatter energy density is from photons left over from the

big bang. These photons make up the cosmic background radiation. For photon energy density, $\rho_E R^4$ remains constant.² The additional factor of R arises because the photons lose energy due to the cosmological redshift. This means that while the photons' energy was dominant early in the history of the universe, the redshift has so reduced their energy that they now are negligible compared to the ordinary mass density. We can ignore the photons when we describe the evolution of the universe today.

For the relativistic equation to take its most complete form, it will be necessary to include the cosmological constant term Λ . However, to begin, we will consider the simplest universe models that are determined by a Robertson–Walker metric, obey the conservation of mass-energy, contain some specified total mass-energy density, and have a cosmological constant equal to zero. These are referred to as the **standard models**, and while they are not the only possible cosmological models, they are those that follow from our minimal set of assumptions; specifically, a universe in which only standard gravity operates and the cosmological principle holds. These models all decelerate, since gravity acts to pull matter together and to slow down the expansion. We know from observation that the expansion factor is increasing at the present time, that is, \dot{R} must be positive; we have also just argued that $\ddot{R} < 0$ for the standard models, that is, the universe decelerates. From this, we can conclude for these models that the function $R(t)$, whatever its form, must have reached $R = 0$ at some time in the past. We may always adjust our cosmic time coordinate such that this occurred at $t = 0$; thus we obtain, as the initial condition of the universe, that

$$R = 0 \quad \text{for } t = 0.$$

Standard models defined

Since the scale factor gives the separation of comoving points at a particular cosmic time, it obviously must be related to the density. But if comoving points have zero separation, then the density must be infinite. This initial state of infinite density is what is meant by the **big bang**. All standard models begin with a bang. The big bang is *not* a point in space; the scale factor is zero at all spatial locations, in accordance with the cosmological principle and the Robertson–Walker metric. In other words, the big bang happened everywhere.

Introducing the big bang

If we briefly consider the big-bang limit of equation (11.10), we see something interesting. The closer the approach to the initial time, the less the geometry of the model matters. Because density ρ is proportional to inverse volume, $\rho R^2 \propto 1/R$. As R shrinks, the first term on the right-hand side of the equation, the mass-density term, dominates over the constant contribution due to the curvature. This is a valuable simplification, since it means that we can describe the earliest stages in the universe without worrying too much about what its true curvature might be. The ultimate fate of the universe, on the other hand, depends very strongly on the spatial curvature, for largely the same reason. As

²See Chapter 12.

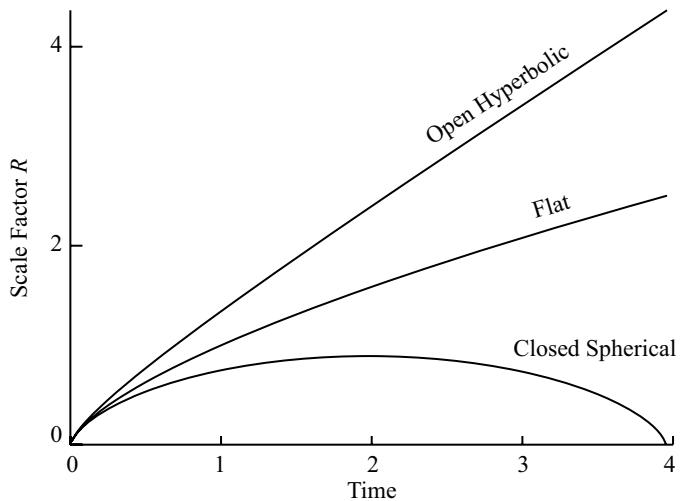


Fig. 11.3 The behavior of the scale factor for the three different geometries of the standard models. All begin with a big bang. The $k = +1$ spherical universe expands to a maximum size, then contracts to a big crunch. While both the flat ($k = 0$) and hyperbolic ($k = -1$) universes expand forever, the hyperbolic universe expands at a faster rate than does the flat universe.

R becomes large the mass density decreases, becoming, in the case that $k \neq 0$, increasingly less important in comparison to the curvature k . Our earlier discussion of the Newtonian model and its relativistic equivalent, the Friedmann equation, also suggested that the fate of the universe is related to its energy at infinity, a role played by the spatial curvature in the standard models. Our energy arguments enable us to predict the future behavior of the models corresponding to each of these geometries: the flat ($k = 0$) and hyperbolic ($k = -1$) models will expand forever, while the spherical model ($k = +1$) will recollapse into something conventionally called the **big crunch**. The spherical model both begins and ends with a bang; it recollapses back to $R = 0$. The flat and hyperbolic models, by virtue of their endless expansion, end not with a bang, but a whimper.

Hubble's law and the scale factor

The Friedmann equation governs the evolution of the scale factor $R(t)$ for the case of a universe described by the Robertson–Walker metric; that is, a universe which is isotropic and homogeneous. A solution to this equation, for a given choice of curvature k and density ρ , is a model of the universe. The scale factor itself is not directly observable; the observables are various quantities derived from it. A model specifies the time behavior of the scale factor, from which observable quantities can be computed. The predictions of the model can then be compared with actual measurements, to evaluate how well that model fits the data. Hence we must obtain theoretical expressions for the observable parameters in terms of those quantities that are specified by the model itself, such as the scale factor, the curvature constant, and any cosmological constant. One of the most important observables is, of course, the Hubble constant H , so let us first investigate how the Hubble constant is related to the scale factor.

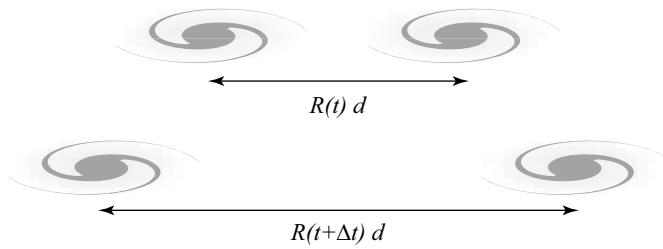


Fig. 11.4 Two galaxies are separated by a comoving distance d . At time t their physical separation is $R(t)d$. At a time Δt later their separation is $R(t + \Delta t)d$. The scale factor increases and the physical distance increases along with it, but the comoving distance remains the same, by definition.

The Hubble law itself can be derived directly from the Robertson-Walker metric. Although a rigorous proof requires some basic calculus, a simplified demonstration can be visualized from Figure 11.4. At a time t , two galaxies are separated by the *comoving* distance d , corresponding to a *physical* distance $\ell(t) = R(t)d$. Since the change in separation of these “test” galaxies is a consequence only of the change in the scale factor, we know that at some later time $t + \Delta t$, their physical separation is $R(t + \Delta t)d$, or, approximately, $(R + \Delta R)d$. (Recall that the comoving distance does not change, by definition.) Thus after an interval of time Δt , the separation of the galaxies has changed by a distance $(\Delta R)d$; dividing the change in distance by the time interval gives a velocity, $v = \dot{R}d$. (Compare equation (11.3).) Since the initial separation was $\ell = Rd$, comparison with the theoretical Hubble law $v = H\ell$ leads to the conclusion that

$$H = \frac{v}{\ell} = \frac{\dot{R}}{R}. \quad (11.11)$$

Hubble constant defined

Because R is a function only of time, its rate of change also depends only on time, and thus the Hubble constant is itself a function of cosmic time alone. Of course, this means it is not truly a *constant*. Homogeneity requires only that it be the same at every point in space at a given instant of cosmic time. Referring to it as a constant derives from its mathematical role as the “constant of proportionality” in the Hubble law, and so the term “Hubble constant” persists. The general symbol for the Hubble constant is H , without a subscript. Its value at the present time, that is, “now,” is denoted H_0 , pronounced “H-naught.” This is the value we could determine through measurements of galaxies that are near to us. Note that with the definition of H it is now possible to rewrite the Friedmann equation (11.10) as a time-evolution equation for the Hubble expansion term, namely

$$\left(\frac{\dot{R}}{R}\right)^2 = H^2 = \frac{8\pi G\rho}{3} - \frac{kc^2}{R^2}. \quad (11.12)$$

This form of the Friedmann equation is particularly useful because it replaces the somewhat theoretical scale factor R with measurable physical quantities like H , density ρ , and spatial geometry.

Observing the standard models

The average mass density of the universe is a potentially measurable quantity. All that is necessary, in principle, is to add up all the mass seen within a large, representative volume of space around us. Needless to say, this is easier said than done. The most obvious stumbling block is that we can detect only luminous matter; the presence of any nonluminous, or “dark,” matter can be inferred only indirectly. Furthermore, it is not easy to find a representative volume. We require a volume large enough to be regarded as homogeneous; but how large would a truly homogeneous volume be? Galaxies cluster, and clusters themselves cluster, complicating our determination of the average matter density, as well as any local deviations from it. Finally, just seeing a galaxy does not tell us its mass. We must somehow deduce its mass from its brightness, or its motions, or from its interactions with its neighbors.

We will discuss the measurement of mass in the universe in more detail in later chapters. The important point is that the average density of the universe has great cosmological significance, making an accurate measurement of this quantity especially important. To show this, simply return to the Friedmann equation in the form of equation (11.12) and rearrange the equation to solve for the curvature in terms of current values of density and the Hubble parameter,

$$\frac{kc^2}{R_0^2} = H_0^2 \left(\frac{8\pi G\rho_0}{3H_0^2} - 1 \right). \quad (11.13)$$

Since we want the terms to be evaluated at time “now,” we have included the “naught” subscripts. Therefore, in the absence of a cosmological constant, we could determine whether the curvature is positive, negative, or zero by measuring the present values of both the Hubble constant and the average density. Through comparisons with observations, we can determine which of the three geometries the actual universe most resembles.

We may draw some interesting conclusions from a careful examination of equation (11.13). First, suppose the universe were empty, that is, $\rho = 0$. Then the curvature (the left-hand side) must be negative, and hence space would have hyperbolic geometry if there were no matter at all. A flat or spherical geometry can occur only in the presence of matter. If we set $k = 0$, corresponding to the spatially flat universe, then

$$\frac{8\pi G\rho_0}{3H_0^2} \equiv \Omega_M = 1, \quad (11.14)$$

where we have defined a new term, the **matter density parameter** Ω_M . In the flat universe the matter density must equal a very specific

critical density, denoted by ρ_c and defined by

$$\rho_c = \frac{3H_0^2}{8\pi G}. \quad (11.15)$$

With this definition we can write the general relationship

$$\Omega_M = \frac{\rho_0}{\rho_c}. \quad (11.16)$$

For a Hubble constant of $100 \text{ km s}^{-1} \text{ Mpc}^{-1}$, the value of this critical density is $\rho_c \approx 2 \times 10^{-26} \text{ kg m}^{-3}$. This corresponds to approximately 10 hydrogen atoms per cubic meter of space, a quantity that is not particularly dense by Earthly standards. Since the critical density scales with H^2 , a Hubble constant of $50 \text{ km s}^{-1} \text{ Mpc}^{-1}$ produces a value one fourth as large, and our current best value for H_0 of $72 \text{ km s}^{-1} \text{ Mpc}^{-1}$ yields $\rho_c \approx 10^{-26} \text{ kg m}^{-3}$.

That the density and the Hubble constant are so intertwined should not be surprising. The Hubble constant is a measure of the velocity of the expansion of the universe; it is this expansion that must be overcome by the gravitational force due to the matter in the universe. The critical universe is precisely balanced; within the standard models a given critical mass density uniquely determines the corresponding value of the Hubble constant, a value that we might call the “critical Hubble constant.” Despite the difficulties in measuring distances, the Hubble constant is the easier of the two quantities to measure, so we generally speak of a critical density implied by the measurement of H_0 , rather than a critical Hubble value implied by a measured density.

Equation (11.13) shows that the value of Ω_M is intimately linked to the geometry of the universe. The value $\Omega_M = 1$ constitutes the boundary between the open, hyperbolic universe and the closed, finite, spherical universe. In the standard models a spherical universe contains sufficient mass-energy for gravity to overwhelm its expansion and cause recollapse, whereas a hyperbolic universe has too little mass-energy to allow gravity to overcome its expansion. The critical universe is exactly balanced; there is sufficient mass-energy to halt the expansion, but only after infinite cosmic time has elapsed. Thus the density of the universe directly determines the geometry of space-time in the standard models. We define the Ω parameter in terms of the *present* density of the universe, although we can also see that for the spherical and hyperbolic universes, the density parameter Ω changes with time because the term $H^2 R^2$, upon which Ω depends, is a function of cosmic time. (An empty universe, with $\Omega = 0$, is an exception; in this case Ω does not change with time.) In the flat universe $\Omega = 1$ for all times.

Using the Ω_M parameter, we can write the Friedmann equation (11.10) in the form

$$\dot{R}^2 = H_0^2 \Omega_M \left(\frac{R_0}{R} \right) - kc^2. \quad (11.17)$$

The value of this form is that the evolution of the universe is written in terms of measurable quantities, namely the Hubble constant and the density parameter.

The geometry and the density of the universe are related in the standard models

The deceleration parameter defined

We have stated that gravity slows the rate of expansion, and to relate this deceleration to observations we define a new parameter, the **deceleration parameter**. The rate of change of the scale factor, \dot{R} , specifies the expansion rate of the universe. The rate of change of the rate of change of the scale factor, conventionally symbolized by \ddot{R} , denotes how the expansion itself changes in time, and this gives us the acceleration of the universe. The negative of the acceleration is, as usual, the deceleration, so we might define this new parameter as simply $-\ddot{R}$; but it is more convenient to define it as a dimensionless quantity, to make it independent of whatever way we may have chosen to set the specific value of R . The standard definition is given by the formula

$$q = -\frac{\ddot{R}}{RH^2}, \quad (11.18)$$

where the inverse factors of R and H perform the role of making q dimensionless. Physically, the deceleration parameter tells us whether the expansion rate of the universe is increasing or decreasing. The name and definition of this parameter suggest a prejudice that the expansion of the universe is slowing. This is the case for many models, because the attractive force of gravity always acts to pull together the matter of the universe, slowing the expansion. Even so, the deceleration parameter can indicate either a deceleration or an acceleration of the expansion by its sign; it is positive for a deceleration and negative for an acceleration. Its value at the present time is denoted by a zero subscript, that is, q_0 (“q-naught”). All standard models are characterized by $q > 0$; that is, *all* these models decelerate.

By combining various formulae developed above for the standard models, we find that the deceleration parameter is directly related to the density parameter by a very simple formula:

$$q = \frac{1}{2}\Omega_M. \quad (11.19)$$

The age of the universe compared with the Hubble time

Thus for the standard models, the specification of a value of q also determines the geometry of space, and hence the specific model.

At this point, the deceleration parameter may still seem like an abstract concept, but it can immediately tell us something about the difference between the actual age of the universe and the Hubble time. If the cosmic expansion is slowing down, then the Hubble constant we measure today will be smaller than it was previously. The Hubble time will thus overestimate the age of the universe. As an analogy, suppose that you drove on a long car trip from point A to point B, a distance of 100 miles, while slowly reducing your speed. If your speed was 20 miles per hour when you arrived at point B, then you would overestimate the length of your trip if you merely divided 100 by 20 to obtain an elapsed time of 5 hours. Conversely, if you accelerated during your trip, then you would underestimate the time of travel if you divided the total distance traveled by your instantaneous speed upon your arrival at point B. In the same way, in a decelerating universe with $q_0 > 0$, the age of

the universe will be *less than* the Hubble time, because at earlier times it was expanding at a faster rate, whereas a universe that has always been accelerating, that is, $q_0 < 0$ for all time, will have an age that is *greater than* the Hubble time. A universe that expands at a constant rate, $q_0 = 0$, has an age equal to the Hubble time.

From these results we have complete limits on the standard models in terms of observables. If the density of the universe is greater than the critical density, the model is a **closed universe** with $q > 1/2$. In such a universe, the curvature constant is $k = 1$, and the energy at infinity is negative. Sufficient matter is present to halt the expansion eventually, at which point the universe begins to contract. The universe then shrinks to ever smaller size, ending in a big crunch at some finite time in the future. This universe has a spherical spatial geometry; it is finite in both space and time. If the density is precisely equal to the critical value, we obtain a **flat universe** with $q = 1/2$. The curvature constant is $k = 0$, and the universe expands to infinity; the expansion would come to a halt only after the passage of an infinite interval of time. This is a special case, because its spatial geometry is Euclidean, or flat, and because the parameters describing it can take only very specific values. The density parameter Ω_M is equal to unity at all times. The universe is infinite in both space and time. (In fact, it is infinite in space for *all* times, not just at infinite time.) This very special case is also known as the **Einstein–de Sitter model**.³ If the density is less than the critical value, then $q < 1/2$, $k = -1$, and we have the third standard cosmological model. This model also expands forever, but the expansion never ceases, even after infinite time. The geometry of this **open universe** is hyperbolic, and it is infinite in space and time. The density decreases faster than does the Hubble “constant,” so the density parameter Ω_M approaches zero at large times.

The Friedmann equation (11.10) provides the means to produce a complete cosmological model: we need merely solve for the function $R(t)$, then use it to evaluate the observable parameters. We have already discussed these solutions qualitatively, but of course there exist precise mathematical solutions for $R(t)$. Since the Friedmann equation is a differential equation, its solution, while straightforward, requires calculus. The easiest solution to obtain is that for the flat (Einstein–de Sitter) model. It is relatively simple to show that

$$R(t) = R_0(t/t_0)^{2/3}, \quad (11.20)$$

where t_0 is some nonzero cosmic time (say, the present time) at which the scale factor R has the value R_0 . The behavior of this scale factor is illustrated in Figures 11.3 and 11.5.

The solutions for other models can be understood in comparison to the flat model. For example, the scale factor R in the closed spherical model

The three standard models: closed, flat, and open

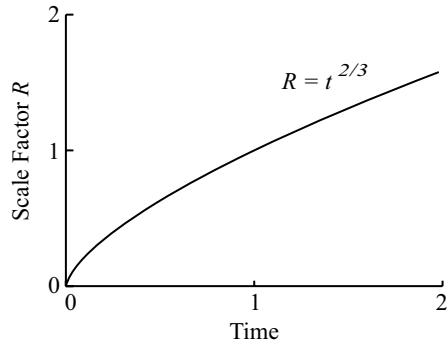


Fig. 11.5 The scale factor R as a function of time for the Einstein–de Sitter standard model. This geometrically flat model begins with a bang at time $t = 0$.

³Do not confuse the Einstein–de Sitter model with either the Einstein static model, or with the empty and expanding de Sitter model. The Einstein–de Sitter model is another name for the flat, matter-filled, standard universe. Obviously, Einstein and de Sitter between them made many early contributions to cosmology.

Standard Friedmann models

Model	Geometry	k	Ω_M	q_0	Age	Fate
Closed Einstein-de Sitter	Spherical	+1	> 1	> 1/2	$t_0 < \frac{2}{3}t_H$	Big crunch
	Flat	0	= 1	= 1/2	$t_0 = \frac{2}{3}t_H$	Expand forever
	Hyperbolic	-1	< 1	< 1/2	$\frac{2}{3}t_H < t_0 < t_H$	Expand forever

increases less rapidly with time, reaches a maximum, and then falls back to zero. In the open hyperbolic model, R increases more rapidly than $t^{2/3}$. For the hyperbolic model with $k = -1$, when t becomes very large, the solution for R is nearly proportional to time. At this point the mutual gravitational attraction of all the mass in the universe is so weak as to have almost no effect upon the expansion of space.

From the explicit formula for the scale factor, the Hubble constant can be computed as a function of time. For the Einstein-de Sitter universe, this is given by the simple result $H = 2/(3t)$; as in all standard models, the Hubble constant decreases with increasing cosmic time. From this expression, it follows that the true age of the critical universe is $2/3$ the Hubble time. Specifically, if the Hubble constant were presently $50 \text{ km s}^{-1} \text{ Mpc}^{-1}$, the corresponding Hubble time would be 20 billion years, and the actual age of the flat universe would be $2/3 \times 20 = 13.3$ billion years. The Einstein-de Sitter solution provides an important dividing line for the standard models. Denser, closed models will have *smaller* ages than that of the Einstein-de Sitter model; open models will have *greater* ages than this flat model. If, for example, we could demonstrate conclusively that some star cluster was older than 13.3 billion years, this datum would rule out the flat and closed standard models unless the Hubble constant proved to be less than $50 \text{ km s}^{-1} \text{ Mpc}^{-1}$.

Models with a cosmological constant

The standard models are not the only possibilities, even within the restriction of homogeneous, isotropic universes. Einstein originally attempted to find a cosmological solution that was both static and spatially finite, with no boundary. He quickly discovered that such a universe cannot *remain* static; gravity's inexorable pull forces such a model to evolve. Rather than accept an evolving universe, however, Einstein added a term, conventionally symbolized by Λ and called the **cosmological constant**, to his equations of general relativity. This quantity, unlike the Hubble "constant," is a true constant; its value never changes for a given cosmological model. It acts over long distances; in Einstein's original formulation, it provides a repulsive force that counters gravity. In the Einstein static model, the effect of this parameter on the scale of

the Earth is immeasurably small, yet on cosmic scales it just balances the tendency for matter to pull the universe toward collapse. Of course, Einstein's attempt to make a static universe ultimately failed, and this led him to recant on the cosmological constant. But the genie could not be put back into the bottle quite that easily. The cosmological constant can be used to create interesting models of the universe, and cosmologists have done so. More significantly, evidence has recently emerged for the existence of a nonzero cosmological constant in the physical universe, due to a source known as *dark energy*.

We thus expand our study of cosmological models by modifying the Einstein equations to include the cosmological term. The relativistic equivalent of Newton's gravitational acceleration equation (11.4) becomes

$$\ddot{R} = -\frac{4G\pi}{3}\rho R + \frac{\Lambda R}{3}. \quad (11.21)$$

The first term on the right is just the familiar gravitational term.⁴ The new term on the right can be interpreted as the acceleration associated with the cosmological constant. With a positive sign, Λ provides a force opposite to that of gravity, that is, a repulsive force, which is just what Einstein needed to make his model static. In principle, the cosmological constant need not be repulsive; it could also be negative, that is, attractive. If the cosmological constant provides another attractive force, it supplements gravity. We shall be primarily concerned, however, with its unique role in providing repulsion.

An important property of the cosmological constant arises from the dependence of the Λ -force upon the scale factor. Whereas the gravitational force of the mass in the universe drops off as the inverse square of R , the Λ -force *increases* with R . This means that when the universe becomes large enough, the Λ -force will inevitably dominate over the gravitational force. As one consequence of this, if Λ is negative and hence augments gravity, the universe will slow down and recollapse no matter how rapid its present expansion or how small Ω_M might be. Hence it is possible to have a recollapsing universe that does not have a spherical geometry.

The presence of Λ means that we must return to the Friedmann equation and recompute the values of the Hubble constant, the critical density, and the deceleration parameter in light of this new term. So far we have introduced Ω_M as the ratio of the matter density to the critical density; in the standard models only ordinary matter controls the evolution of the universe. Now we add the Λ term, and it too can be characterized in terms of an Ω parameter. Returning to the Friedmann equation, write

$$\dot{R}^2 = \frac{8}{3}\pi G\rho R^2 + \frac{\Lambda R^2}{3} - kc^2. \quad (11.22)$$

The Friedmann equation with Λ

⁴We have again assumed here that any energy density other than ordinary mass is negligible. Specifically, we ignore the contributions from the cosmic background radiation, an approximation that is valid for most of the history of the universe.

If we combine this with equation (11.21) and use the definitions of our various cosmological parameters, we find the following relationship:

$$\left(\frac{\dot{R}}{R}\right)^2 = H^2 = \frac{8\pi G\rho}{3} + \frac{\Lambda}{3} - \frac{kc^2}{R^2}. \quad (11.23)$$

We can rewrite this for the present time t_0 in the form

$$\frac{kc^2}{R_0^2} = H_0^2 (\Omega_M + \Omega_\Lambda - 1). \quad (11.24)$$

This form of the equation leads to

$$1 = \Omega_M + \Omega_\Lambda + \Omega_k, \quad (11.25)$$

where we define the **lambda density parameter** as

$$\Omega_\Lambda = \Lambda/3H_0^2 \quad (11.26)$$

and the *geometry parameter*

$$\Omega_k = -\frac{kc^2}{R_0^2 H_0^2}. \quad (11.27)$$

An important consequence of the inclusion of Λ that is apparent from these equations is that the simple relationship characterizing the standard models becomes more complicated. The matter density Ω_M no longer uniquely determines the geometry of the model. Rather, it is the sum of distinct Ω parameters that must add to unity. The virtue of writing the equation in this form now becomes a bit clearer. By putting Λ into the form of Ω_Λ we have a simple measure of the relative significance of Λ compared to ordinary matter. We can thus describe the evolution of the universe through these measurable parameters by writing the Friedmann equation in the form

$$\dot{R}^2 = H_0^2 \left[\Omega_M \left(\frac{R_0}{R} \right) + \Omega_\Lambda \left(\frac{R}{R_0} \right)^2 + \Omega_k \right]. \quad (11.28)$$

Now let us examine some of the types of models that are possible when Λ is present. Historically the first such model was Einstein's static, spherical universe. We can use equation (11.21) to compute the special value of Λ that Einstein calculated for his universe. Setting \ddot{R} to zero (zero acceleration, no net force) we find that the Einstein critical value of Λ is

$$\Lambda_c = 4\pi G\rho. \quad (11.29)$$

From equation (11.22), we further obtain $kc^2/R^2 = \Lambda_c$. Thus in the Einstein static universe, the cosmological constant has a positive value that is determined by the average density of the universe. We can see that it is also equal to the spatial curvature, which is positive; thus this model corresponds to a spherical geometry.

A very different cosmology, which contains a cosmological constant but is devoid of matter, is the eponymous **de Sitter model**. The de Sitter universe has flat spatial geometry ($\Omega_k = 0$), zero density ($\Omega_M = 0$), and a positive cosmological constant ($\Omega_\Lambda = 1$). From equation (11.23), we find that the Hubble constant is truly a constant, and

$$H = \sqrt{\Lambda/3}. \quad (11.30)$$

Since H is a genuine constant in this case, the time dependence of the scale factor can be determined by elementary calculus from the equation $H = \dot{R}/R$, with the result that

$$R(t) = R_0 e^{Ht/t_0}, \quad (11.31)$$

where R_0 is the value of the scale factor at time t_0 .⁵ The behavior of this *exponential* solution is dramatic; as Figure 11.6 indicates, the increase in the scale factor with time is very rapid. From the definition of the deceleration parameter, equation (11.18), we can determine that $q = -1$; that is, this universe is accelerating.

A feature of the exponential curve is that its appearance is the same everywhere. That is, any section of the curve can be overlaid on another section, with just a change in scale. As one consequence of this, the de Sitter universe is infinitely old and never goes through a big bang. Running backwards in time, $R(t)$ shrinks to smaller and smaller values, but never reaches zero. All this may seem so strange that it could not possibly have anything to do with the physical universe, but the de Sitter universe describes the behavior of several important models. One is a possible *inflationary epoch* in the early universe, an epoch that we shall describe in more detail in Chapter 16. Another example is the late evolution of any open, expanding universe with matter and a positive cosmological constant. As the universe expands, ρ drops toward zero and $\Omega_M \rightarrow 0$. As it does so, the model behaves more and more like a de Sitter universe.

Another example of a historically interesting, exponentially expanding model, albeit a rather more extreme one, is the **steady state model**. The steady state model obeys an idealization known as the *perfect cosmological principle*.⁶ The perfect cosmological principle holds that not only is every point in space representative of the universe as a whole, but each point in time is representative of the entire history of the universe. In other words, the steady state universe is isotropic and homogeneous in time as well as in space. As in the de Sitter universe, the Hubble constant is truly a constant, the same for all times, and again this yields an exponential expansion. This model expands without a big bang, and continues to expand in the same manner forever. The universe has always existed, and always will exist. An important difference between

The de Sitter universe

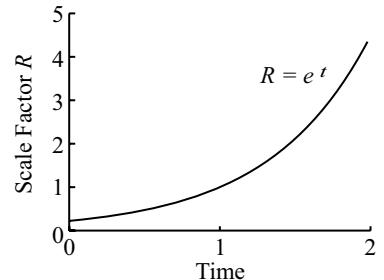


Fig. 11.6 The scale factor R as a function of time for a de Sitter universe. The exponential curve of the de Sitter model never goes to $R = 0$, so there is no big bang; this model is infinitely old.

The de Sitter model describes the distant future of a $\Lambda > 0$ accelerating model.

The expanding steady state model

⁵For those unfamiliar with it, e represents a special number that occurs frequently in mathematics. It is the base of the so-called natural logarithms, and it has many other interesting properties. Like π , it is a transcendental number; it happens to be given by $e \approx 2.718\dots$

⁶The perfect cosmological principle is discussed in Chapter 6.

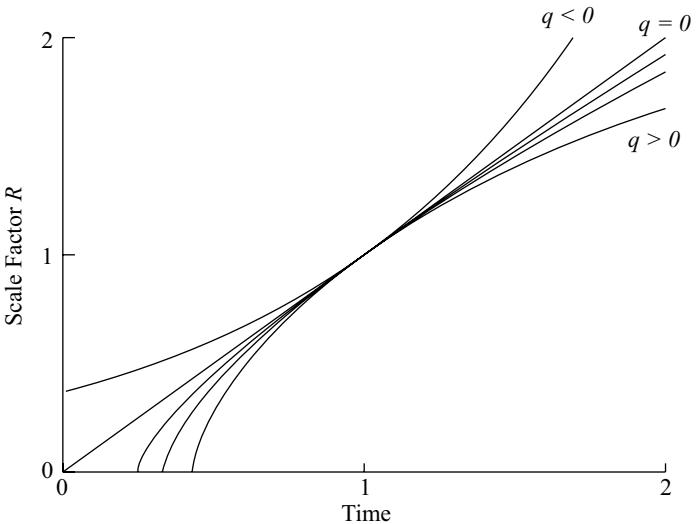


Fig. 11.7 The behavior of the scale factor for a variety of models. All are constrained to pass through the present time, arbitrarily set to $t_0 = 1$, with the same slope. The value of the deceleration parameter determines both the model's future and the age of the universe. The larger the deceleration, the shorter the time back to the big bang. The exponentially expanding de Sitter model never intersects $R = 0$.

the de Sitter model and the steady state model is that the latter may contain matter. But if this universe is expanding, should not the density decrease as the third power of the expansion factor? If the model is to adhere to the perfect cosmological principle, this cannot occur. The steady state cosmology thus requires the existence of a “creation field” that creates mass-energy at precisely the correct rate to balance the expansion. The creation field is also responsible for the acceleration, since the ordinary mass-energy could produce only a deceleration. In other words, the creation field generates the cosmological constant Λ , with a value given by equation (11.30).

The steady state model demands the introduction of new physics, the creation field, beyond the equations of general relativity and other standard laws of physics. That alone need not rule this model out; but the steady state cosmology has been emphatically rejected by observations. For one thing, the perfect cosmological principle requires that the average appearance of the universe remain unchanged for all time, yet observational evidence shows that the universe has evolved considerably over the past several billion years. The universe during the time of the quasars was a very different environment from what we see around us at the present epoch. Another test of the steady state cosmology is to measure the deceleration parameter q . In an exponentially expanding model, $q = -1$.⁷ Although measuring q is difficult, the data, while still inconclusive in obtaining a precise value, do not seem to be at all consistent with the value $q = -1$. Finally, the most serious blow to the steady state model was the detection of the cosmic background radiation. This provides direct evidence for the hot initial state we call the

The cosmic background radiation provides evidence for the big bang

⁷To show this, combine the definition of q , equation (11.18), with the solution for R in equation (11.31).

big bang.⁸ The steady state model was never particularly popular with cosmologists or astronomers, although it was once advocated by a vocal minority as an alternative to the big bang. Most astronomers regard the steady state cosmology as an historically interesting model that is in conflict with the observations.

The de Sitter universe, the Einstein static universe, and the steady state model are all quite different from the standard models, and all three are ruled out by observations as descriptions of the present universe. However, they are by no means the only possible models that contain a cosmological constant. The Λ term can be used to derive many interesting alternatives. A dramatic example is obtained by the addition of a *negative* Λ term. Because a negative Λ augments gravity, and becomes stronger with increasing scale factor, it will cause any geometry ultimately to recollapse, even the flat and hyperbolic models. With the right balancing of terms, such models can exist for an indefinite period, but the return to the big crunch is inevitable. Adding a negative Λ to the spherical model, which was already fated to recollapse, causes the end to come just that much sooner.

More interesting, and, as it turns out, more realistic possibilities are created by the addition of a positive (repulsive) Λ term to otherwise standard models. A positive Λ will not change the ultimate fate of either of the open models, whether flat or hyperbolic, but it will change their behavior as they expand. As the Λ -force begins to dominate over gravity, the universe will start to accelerate, and q will become less than zero. The acceleration will increase until the models are expanding exponentially. Eventually, they will behave like empty de Sitter universes, as the gravitational attraction of the increasingly low matter density becomes utterly irrelevant.

Adding a positive Λ to the spherical standard model produces additional possibilities. The fate of this spherical model depends on the exact value of the cosmological constant. Einstein introduced his Λ term to provide a balance with gravity; thus the Einstein critical value Λ_c determines whether a spherical model with a cosmological constant will recollapse or will expand forever. If $\Lambda < \Lambda_c$, models with spherical geometry will recollapse, although the closer Λ is to the critical value, the longer the model lasts before the big crunch. On the other hand, a spherical model with $\Lambda > \Lambda_c$ will be prevented from recollapsing by the Λ force. One such model is the **Lemaître model**, devised by Georges Lemaître. If the cosmological constant is barely larger than Λ_c , the universe may spend a very long time with gravity and the Λ force in near balance, in a lengthy, nearly static, hovering period. After this hovering interval, the universe begins to expand again, now at an accelerating rate. There would be interesting observable consequences if we lived in an accelerating Lemaître universe. In the Lemaître model, R remained nearly constant for a long period in the past. Because of this, there could be many objects at the special redshift associated with the hovering in-

All models with $\Lambda < 0$ recollapse

A positive Λ can produce an accelerating universe

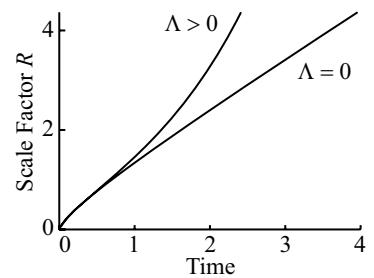


Fig. 11.8 When a positive Λ term is added to the hyperbolic standard model ($\Lambda = 0$), R expands more rapidly, and soon accelerates with $q < 0$.

⁸See Chapter 14.

An interesting “hovering” universe

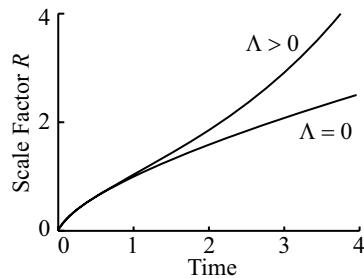


Fig. 11.9 When a positive Λ term is added to the flat standard model ($\Lambda = 0$), R eventually begins accelerating with $q < 0$, although at a later time compared with the hyperbolic model.

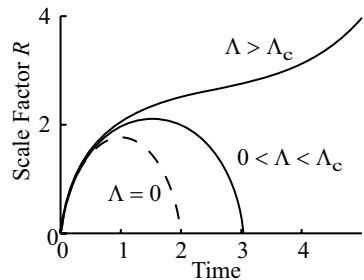


Fig. 11.10 The Lemaître model has spherical geometry and a Λ value slightly greater than the Einstein critical value Λ_c . This model features a hovering period, during which the scale factor remains nearly constant over a lengthy time interval. Following the hovering period, expansion continues at an accelerating rate with $q < 0$.

terval. This was of some interest in the 1970s when there appeared to be an excess of quasars at a redshift of approximately $z = 2$. However, the excess turned out to be largely a selection effect; the technique used to find quasars favored those that had a redshift close to two. Since then, better observations have shown that there was a great age of quasars when the universe was younger, but they are distributed over a range of redshifts. Nothing about the number of quasars at any particular redshift now indicates that there was ever a hovering period. Another observable consequence is that in the Lemaître model the age of the universe can be *much* greater than the Hubble time. This is another reason why this particular model was once of some interest to cosmologists. Hubble’s initial measurements gave a very large value of H_0 , which was only slowly revised downward. These early observations indicated a Hubble time close to two billion years, much less than the age of the Earth! In the Lemaître model, the Hubble age would provide only a rough estimate of the time back to the hovering period; the universe could be considerably older than that. As the Hubble time and the ages of the constituents of the universe came into better agreement, however, the Lemaître model lost its brief popularity.

One final consequence of this model is particularly interesting. In any closed spherical geometry, it is possible for light to travel all the way around the cosmos. For example, we could look far into space and see our own Milky Way forming. In the standard spherical model, the transit time for light around the universe is the same as the entire lifetime of the universe, so that by the time we can “see ourselves,” we are caught in the big crunch. This limitation is overcome in the Lemaître model, with its static period. In this case, the universe might be sufficiently old for some photons to have had time to travel completely around the universe. The light from a distant quasar exactly halfway around the universe could arrive from two opposite directions at the same time; we would thus observe the same quasar in opposite directions in the sky. Searches were made for such “mirrored” quasars, but none were detected.

A historically important aspect of the Lemaître model is that it begins from a state of large or infinite density. Lemaître seems to have been the first to take seriously such an initial state, which he called the “primeval atom.” It can justifiably be said that Lemaître paved the way for acceptance of the later big bang models. We shall return to this historical thread in Chapter 12.

In the final analysis, how should we regard the cosmological constant term Λ ? Is it just a free parameter, a “fudge factor,” thrown in to adjust the models as needed, or does it have a serious role to play in cosmology? It is possible, in principle, to measure its value, though as is usual in observational cosmology, such measurements are difficult. However, recent observations have opened up the possibility that our universe does contain a nonzero, positive Λ . The evidence continues to grow that this mysterious quantity may also have played a role in the early moments of the universe, and that it may come to dominate the

Cosmological models

Model	Geometry	Λ	q	Fate
Einstein	Spherical	Λ_c	0	Unstable; collapse or expand
de Sitter	Flat	> 0	-1	Exponential expansion
Steady state	Flat	> 0	-1	Exponential expansion
Lemaître	Spherical	$> \Lambda_c$	< 0 after hover	Expand, hover, expand
Closed	Spherical	0	$> 1/2$	Big crunch
Einstein-de Sitter	Flat	0	1/2	Expand forever
Open	Hyperbolic	0	$0 < q < 1/2$	Expand forever
Negative Λ	Any	< 0	> 0	Big crunch

cosmos in the far distant future. It may happen that, due to processes from particle physics, Λ is different at different times in the history of the universe. In such circumstances, Λ could change as the universe undergoes transitions from one state to another. Such a universe could not be approximated by a single model over its existence, but would move from one model to another. An understanding of the Λ models is essential to comprehending the frontiers of cosmology.

Chapter Summary

A model of the universe is a mathematical description of how the scale factor $R(t)$ evolves with time. As a first approximation, we consider the analogy of the Newtonian ball of self-gravitating particles. Gravity acts to try to pull the ball together. If the ball is expanding with sufficient velocity, it can resist this collapse. We obtain a simple equation to describe the evolution of this Newtonian ball. One of the most important consequences of this analysis is the realization that gravity permits three possibilities: the ball could expand forever, it could stop expanding at infinite time, or it could stop expanding at some finite point in time and recollapse.

Remarkably, the fully general-relativistic solution for a universe consisting of smoothly distributed matter has the same form as the Newtonian solution, although general relativity adds an overall space curvature, symbolized by the constant k . Three geometries are possible: the familiar flat geometry with zero curvature, spherical geometry with positive curvature, and hyperbolic geometry with negative curvature. The general relativistic equation that describes the evolution of the universe under

the influence of its self-gravity is called the Friedmann equation. Models in which only gravity operates (that is, no cosmological constant is present) and mass-energy is conserved are called standard models. The three possible fates of the universe correspond to the three basic geometries: the hyperbolic universe expands forever; the flat universe expands forever, but ever more slowly; while the spherical universe reverses its expansion and collapses in a big crunch.

Adding a nonzero cosmological constant, Λ , provides a number of new possible models. The Λ term acts as an additional force, either attractive (negative lambda) or repulsive (positive lambda). Rather than decreasing in strength with distance like gravity, the Λ force increases with distance, so any nonzero cosmological constant will ultimately dominate the evolution of the universe.

A cosmological model is characterized by its matter content, the presence or absence of Λ , and the spatial curvature k . These values can be conveniently parameterized in terms of a critical density parameter Ω , which is the ratio of the quantity, converted as necessary to units

equivalent to density, to the critical density of the universe. The Friedmann equation can then be written in a particularly simple form in which the sum of the Ω terms

must be equal to 1. These parameters are potentially observable, allowing us to select the best model to describe the actual universe.

Key Term Definitions

Friedmann equation An equation describing the evolution of the cosmological scale factor of the Robertson–Walker metric.

standard model One of the set of big bang cosmological models that are generated with the minimum set of assumptions, namely that the cosmological principle holds and the cosmological constant is zero.

big bang The state of extremely high (classically, infinite) density and temperature from which the universe began expanding.

big crunch The state of extremely high density and temperature into which a closed universe will recollapse in the distant future.

matter density parameter The ratio of the average density in gravitating matter to the critical density, written Ω_M .

critical density That density which just stops the expansion of space, after infinite cosmic time has elapsed. In the standard models, the critical density requires that the spatial geometry be flat.

deceleration parameter A parameter, q , that denotes the rate of change with time of the Hubble constant.

closed universe A standard model with a spherical three-dimensional spatial geometry. Such a universe is finite in both space and time and recollapses.

flat universe A model whose three-dimensional spatial geometry is flat.

Einstein–de Sitter model The flat ($k = 0$), pressureless standard model of the universe.

open universe A model that expands forever and is infinite in space and time, although it begins with a big bang. Sometimes applied strictly to the hyperbolic standard model, although both the hyperbolic and flat models are open in the sense of expanding forever.

cosmological constant A constant introduced into Einstein’s equations of general relativity in order to provide a supplement to gravity. If positive (repulsive) it counteracts gravity, while if negative (attractive) it augments gravity. It can be interpreted physically as an energy density associated with space itself.

lambda density parameter Analogous to the matter density parameter, this term, written Ω_Λ , measures the relative importance of the Λ term compared to the critical value that would correspond to a flat universe.

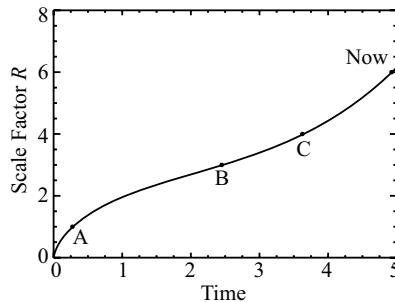
de Sitter model A model of the universe that contains no matter but only a positive cosmological constant. It expands exponentially forever.

steady state model A cosmological model that obeys the perfect cosmological principle. Generally applied to specific models that contain a cosmological constant generated by the regular creation of matter.

Lemaître model The cosmological model developed by Georges Lemaître, which contains a positive cosmological constant, uniform matter density, and spherical spatial geometry.

Review Questions

- (11.1) Using the Friedmann equation along with the definition of q , show that the deceleration parameter q is equal to $1/2$ for a $k = 0$ standard universe. (This requires some algebra.)
- (11.2) Describe the expanding steady state model. Describe an observation that would test the predictions of the perfect cosmological principle. Is the steady state model in conflict with present observations of the universe?
- (11.3) Suppose it were discovered that the Hubble time was 17 billion years, and the oldest stars were 15 billion years old. Among the standard ($\Lambda = 0$) models, which would be acceptable? What possibilities open up if a cosmological term Λ is added to the model?
- (11.4) We have observed quasars with redshifts as large as $z = 4$. How large was the universe then compared to now? A useful quantity is the lookback time, the travel time required for light with a certain cosmological redshift to reach us. The actual value of the lookback time depends on the specific model. For the flat Einstein–de Sitter model it is
- $$t_{\text{lb}} = \frac{2}{3H_0} \left(1 - \frac{1}{(1+z)^{3/2}} \right)$$
- for redshift z . Using this formula, what is the lookback time to the $z = 4$ quasar, if the Hubble time is 20 billion years? If the universe were the closed spherical model, would the lookback time be larger or smaller than that for the flat model?
- (11.5) How does the critical density parameter Ω depend on the Hubble constant? If the universe were found to have a density equal to the critical value for a Hubble constant of $50 \text{ km s}^{-1} \text{ Mpc}^{-1}$, what would Ω be for a Hubble constant of $100 \text{ km s}^{-1} \text{ Mpc}^{-1}$? For $H_0 = 25 \text{ km s}^{-1} \text{ Mpc}^{-1}$?
- (11.6) How would a nonzero cosmological constant change the evolution of each of the three standard Friedmann models? Illustrate with diagrams for $\Lambda < 0$, $0 < \Lambda < \Lambda_c$, and $\Lambda > \Lambda_c$, where Λ_c is the critical cosmological constant for the Einstein static model.
- (11.7) (More challenging.) Demonstrate by direct substitution that the relation $R \propto t^{2/3}$ satisfies the Friedmann equation for the case of the matter-filled flat universe. (This requires a little calculus.)
- (11.8) Briefly describe the history $R(t)$ of a cosmological model you like, giving the value of its defining parameters (e.g. Ω , k , Λ). Explain why you prefer it; your reasons may be philosophical, theoretical, or observational.
- (11.9) Consider the figure showing a Lemaître universe. Labeled on it are several points. Answer these questions: what sign does the deceleration parameter have at the time marked Now? What sign did it have for quasar A? What redshifts do quasars B and C have? What is the lookback time to quasars B and C (in the time units of the plot)? Is the NOW-measured Hubble period greater or less than the actual age of the universe?



This page intentionally left blank

The Early Universe

12

Key Terms:

The universe contains the record of its past the way that sedimentary layers of rock contain the geological record of the earth's past.

Heinz R. Pagels

Approaching the big bang

The recognition that the universe is expanding leads naturally to the question of its origins. From what might the universe have expanded? What might the universe have been like when it was much smaller? What does “smaller” mean for a universe? Within the standard models, a straightforward projection to earlier time leads to the conclusion that the universe was once much denser and more compact than it is today. Indeed, taking this to its ultimate limit, the universe was *infinitely* dense at that cosmic time when the scale factor R was zero. If the universe began with a big bang, what was this event like? Can we learn anything about it today, or is it too far from our experience to try to understand? How can we even begin to think about the universe near its beginning?

Prior to 1965, little was known with certainty about conditions in the early universe, and the case that could be made for any of the big bang models was no more compelling than were arguments for other models. Many astronomers developed models during the era between the two World Wars, but beyond the bare knowledge that distant galaxies were receding, scarcely any data were available at the time. Hence these models tended to be based more upon philosophy than upon data. Many theorists of the time shared an aversion to an explicit beginning for the universe. For example, Arthur Eddington, the astronomer who was one of the first to realize that nuclear processes must power the stars, devised his own model in the 1930s, in which the universe emerged calmly and gradually from an infinitely distant, nearly static initial state. The Eddington model was essentially an Einstein static model with positive cosmological constant that, after an unknown length of time, caused the universe to begin expanding. Through this contrivance, Eddington avoided the question of an initial state in the finite past.

Another astronomer and stellar theorist, E. A. Milne, rejected entirely any cosmological explanation in terms of general relativity. Milne’s model, which he derived in the 1930s, was based upon special relativity.

- nucleosynthesis
- thermal equilibrium
- matter era
- radiation era
- pair production
- quantum mechanics
- Compton wavelength
- graviton
- symmetry
- spontaneous symmetry breaking
- Planck epoch
- unified epoch
- grand unified theory
- inflation
- hadron
- lepton
- quark
- baryon
- baryogenesis
- quark epoch
- gluon
- electroweak interaction
- hadron epoch
- lepton epoch
- nucleosynthesis epoch
- deuterium
- equal density epoch
- structure formation
- recombination
- heat death
- entropy
- second law of thermodynamics
- arrow of time

Early models of the expanding universe

There was no gravity at all on the cosmological scale. He adopted the point of view that the apparent expansion of the universe was simply the result of an infinity of galaxies expanding outward at ever-increasing velocities approaching the speed of light. The outer edge of this ensemble was identified with a sphere, expanding at the speed of light into flat Minkowskian space; within the sphere, increasingly distant galaxies were Lorentz-contracted by just the right amount to fit an infinite number of galaxies within the finite volume of the sphere. Such a universe obeys the cosmological principle, though this may not be immediately obvious. Because the speed of light can never be reached, the view from each galaxy is the same; surrounding galaxies move according to the Hubble law. Milne's model is mathematically equivalent to the empty hyperbolic standard model, to which Milne's name is sometimes now attached. Milne himself recognized the equivalence but he disliked the idea of curved space, preferring his own interpretation.

Eddington's and Milne's models represent interesting, but futile, attempts to explain the data existing at their time in a manner consistent with their philosophical prejudices. We should not immediately dismiss such efforts as foolish or old-fashioned; aesthetical considerations continued to guide many cosmologists throughout the 20th century. When little data are at hand, not much else is available to aid in the construction of models. Better data from space-based, and improved ground-based, telescopes have revolutionized cosmology in the last two decades, making it possible for cosmologists to rely more upon observations and less upon speculation; but all cosmological observations are very difficult, and not always good enough to be of much help. Significant progress has often occurred because some scientists held stubbornly to a particular viewpoint in the face of apparently contradictory data that was later proved to be wrong. Yet cosmologists must always be prepared to give up their preferred models if the weight of data refutes them. It is a fine line to walk, but there is no other option.

Steady state cosmology

The school of philosophically guided cosmology reached a pinnacle in the steady state model, a theory first advanced in 1948. Hermann Bondi and Thomas Gold developed one version of a steady state cosmology, while Fred Hoyle simultaneously worked out another. Bondi and Gold were uncomfortable with general relativity; they based their model more directly upon Mach's principle, and particularly on the perfect cosmological principle. Hoyle, in contrast, developed a relativistic model that maintained a constant density by the introduction of a new physical phenomenon, a *creation field*. Both models required the spontaneous generation of new matter, usually assumed to take the form of hydrogen. Hydrogen atoms were postulated to appear as necessary to balance the expansion; however, the rate of production of new matter was so small, only about 10^{-24} protons per second per cubic centimeter, that it could not possibly be directly observed. Such a minute quantity of matter creation may seem like only a tiny violation of known physical principles, but the creation of *any* amount of matter out of nothing within the physical universe would require a significant modification in

our understanding of physical law. The proponents of the steady state model never provided a physical theory to account for such a creation.

In contrast to the steady state and other eternal models, the big bang models must deal with the evolution of the universe from an initial state. Around 1931, Georges Lemaître became the first to advocate an explicit beginning. His model consisted of a spherical geometry that contained both matter and a positive cosmological constant. It began with a dense initial state from which the universe rapidly expanded, followed by a lengthy hovering period during which the cosmological constant nearly balanced gravity. During this time the universe looked much like an Einstein static universe, and the scale factor changed very slowly. Finally, the cosmological term won out and the universe resumed a rapid and accelerating expansion.

During the 1930s, Lemaître was one of the few scientists to take seriously the concept of a dense initial state. Hubble was an early convert, and some of Hubble's colleagues worked on the theory of an expanding universe; H. P. Robertson's name is attached to the metric he devised to describe such a model, and Richard Tolman developed the thermodynamics of an early universe dominated by the energy of radiation. Apparently, however, neither Robertson nor Tolman was particularly inclined to make the leap of asserting that the model was a valid description of physical reality. It was Lemaître who took the models as representative of reality, and unhesitatingly explored their ultimate consequences. He believed that the universe began with a density comparable to that of an atomic nucleus and then, in a process he likened to radioactive decay, the particles split apart to ever lower densities. Lemaître had little training in quantum physics nor, apparently, much interest in it; he envisioned the beginning of the universe as a "fireworks" of radiation, which, he speculated, might provide the explanation for cosmic rays.

The big bang did have other early proponents, especially among some nuclear physicists. Novel and bold theoretical calculations, aided by improvements in the understanding of nuclear physics, provided new avenues of investigation. George Gamow and his collaborators, especially Ralph Alpher and Robert Herman, performed the first calculations of **nucleosynthesis**, the creation of elements, in the early universe. They postulated that the universe began as pure neutrons, some of which decayed to create protons plus electrons and antineutrinos. All the elements were then built up via neutron capture. To prevent all matter from ending up as helium, they concluded that the early universe must have been hot; that is, a large number of high-energy photons had to have been present for every nucleon. Gamow and Alpher described this model in the so-called " $\alpha\beta\gamma$ " paper, published in 1948 in the *Physical Review* under the names of Alpher, Hans Bethe, and Gamow.¹ Gamow, Alpher, and Herman realized that this radiation would have eventually

Georges Lemaître was the father of the big bang

The big bang and the creation of the elements

¹Bethe's name, which is pronounced "beta," was added for humorous effect. At one point, Gamow even tried to induce Herman to change his name to "Delter."

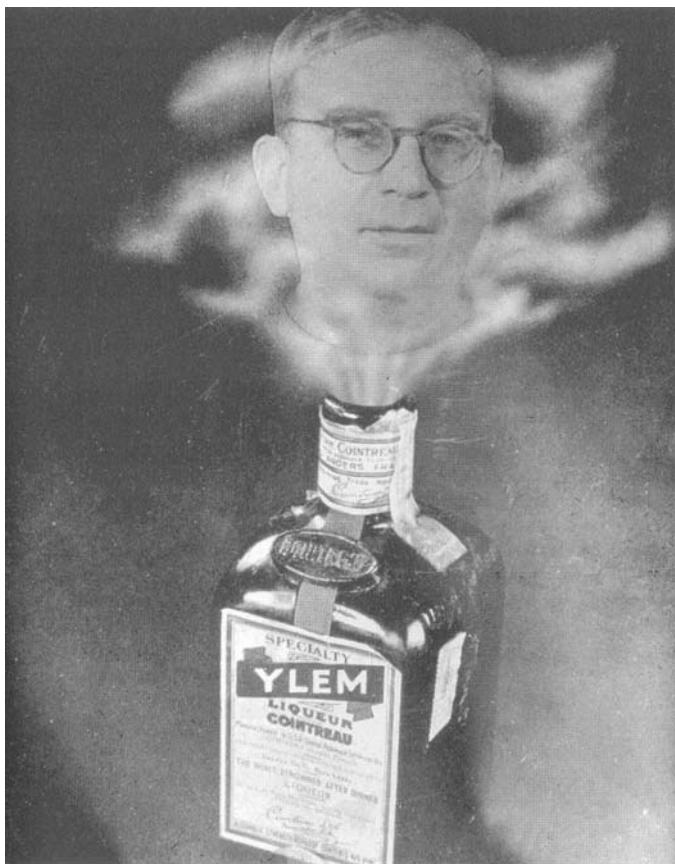


Fig. 12.1 George Gamow (1904–1968). Gamow's calculations of primordial nucleosynthesis led to the first prediction of the existence of the cosmic background radiation. Alpher and Herman created this image of Gamow rising like a genie from a bottle of ylem, his name for the primordial stuff of creation. (Courtesy of Dr. Ralph Alpher.)

A prediction of the cosmic background radiation

escaped from the primordial *ylem*, as Gamow called the hot initial state; hence the relic radiation would still be present in the universe, although greatly redshifted in energy due to the overall expansion. Gamow's original calculations implied a present temperature for this radiation of about 10 K. Alpher and Herman later improved on these calculations and explicitly noted that the present temperature of the relic radiation would be approximately 5 K, not far from the value measured when the radiation was actually observed fourteen years later. This was the first theoretical prediction of the *cosmic background radiation*, a relic of the hot big bang. Unfortunately, although Gamow, Alpher, and Herman's contributions to big bang nucleosynthesis were widely recognized, the prediction of a background of low-temperature radiation throughout the universe was not appreciated. The detection of the cosmic background radiation was serendipitous, and the full importance of this early work was realized only in retrospect.

Alpher and Herman, later working with James Follin, continued to develop the theory of nucleosynthesis; the trio published an important paper in 1953. Their later model was more modern, hypothesizing that the primordial mixture consisted of photons, neutrinos, and both neu-

trons and protons. They succeeded in predicting an abundance of helium of approximately 25% by mass, which agreed quite well with observations of the solar helium abundance. However, they were stymied by the mass gaps, which might be better termed stability gaps, at atomic mass 5 and 8; no stable nuclides exist with those masses. There seemed to be no way to bridge this gap and create carbon from helium. Hence their nuclear progression was halted, and they were unable to explain the origin of elements heavier than helium. As late as the early 1960s, many scientists preferred an alternative explanation, that all elements were formed in the stars. Some of the theory of stellar nucleosynthesis was originally motivated by the steady state model, which at the time was still a viable competitor to the big bang models.

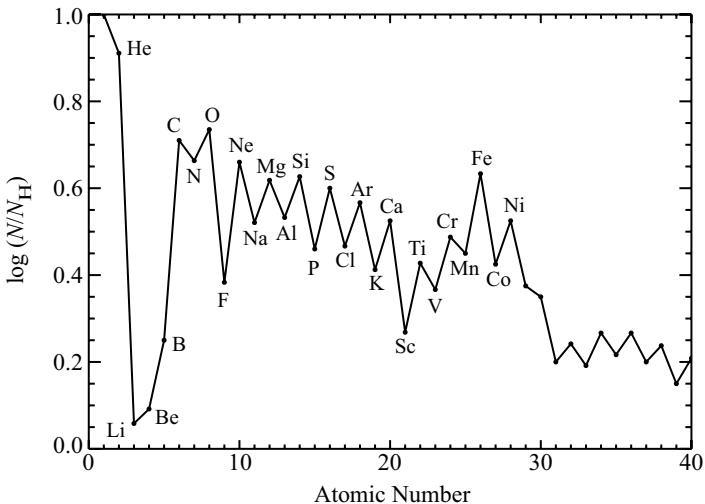
In contrast to big bang models, the steady state model never experiences a hot, dense phase in which nuclear fusion could occur, and therefore it must explain all heavy elements as originating in stellar cores; theoretical work seemed to show that this was possible. Edwin Salpeter demonstrated in 1952 how the stability gaps could be bridged in stars. In 1957, Geoffrey Burbidge, Margaret Burbidge, William Fowler, and Fred Hoyle wrote the definitive paper on the theory of nucleosynthesis of the heavy elements in stars. The stability gap arises from the fact that the fusion of two helium nuclei produces an unstable isotope of beryllium, ^{8}Be , which promptly decays back into two helium nuclei. How do stars jump over this gap in the elements to reach the stable isotopes further along in the Periodic Table? Within stars, the gap is overcome by the so-called *triple-alpha* process, through which helium can be converted to carbon. Although ^{8}Be is unstable, enough of this isotope will be present at the high temperatures and helium densities found in the cores of massive stars that occasionally a nucleus of ^{8}Be captures a helium nucleus, forming a ^{12}C nucleus. The early universe never achieved the high densities and temperatures appropriate to this reaction after helium came into existence; stars are the furnaces in which the heavy elements are forged.

The stars are the source of all the heavier elements of the Periodic Table, from carbon on up. The common isotopes of the elements between helium and carbon (beryllium, lithium, and boron) cannot be generated by ordinary stellar nucleosynthesis, but are produced mainly by reactions involving *cosmic rays*. Cosmic rays are high-energy, relativistic particles, mostly protons, that are ejected from pulsars, supernovae, and other energetic sources. When these particles traverse interstellar gas, some collisions with the gas particles are inevitable. If the cloud has been enriched with carbon and oxygen by earlier generations of stars, a proton will occasionally strike a nucleus of one of these atoms; with so much energy, the proton literally knocks the nucleus apart, creating the light elements. The rarity of these formation processes accounts for the scarcity of these isotopes; they are by far the least abundant of the elements lighter than iron. Still, with one exception (^{7}Li), the formation of these isotopes depends heavily upon prior nucleosynthesis in stars, since carbon and oxygen must be present for the reactions to occur.

Successes and failures of big bang nucleosynthesis

The theory of stellar nucleosynthesis

Fig. 12.2 The relative abundance of the first 40 elements. Hydrogen and helium are by far the most common elements. Most of the helium is created in the big bang, while elements from carbon onward are created in stars. Elements beyond carbon are synthesized by additional fusion processes that occur in the more massive stars; the elements beyond iron are produced in supernova explosions of such stars. Note the deep minimum for the light elements between helium and carbon; those elements are formed neither in the big bang nor in stars, but are produced by nuclear reactions involving the collisions of high-energy free particles with atoms of carbon and oxygen.



The elements are created in both the big bang and in stars

Despite the success of the stellar theory of nucleosynthesis, there remained the problem of explaining the large abundance of helium in the universe. The stars can create helium, of course; the Sun and other main sequence stars obtain their energy by the fusion of hydrogen into helium. Even so, it was extremely difficult to demonstrate how the stars could create *so much* helium. It was already established by the 1950s that the mass of even the oldest of stars consisted of approximately 25% helium, well in line with the prediction of Gamow and his successors, but far more than could be easily accommodated by stellar theory. Not until the acceptance of the hot big bang model did it become clear that this was another instance in which two theories were both right. The big bang creates helium, as well as trace quantities of a few other light isotopes; all others are the products of stars or stellar explosions. In retrospect, this seems like an obvious reconciliation of the mutual difficulties of the two models, but at the time, it was widely believed that it must be one *or* the other. Science is not immune to philosophical prejudices and idiosyncratic blind spots; but over time, the pieces generally fall into place.

The expanding big bang model, in one form or another, gradually became sufficiently widely known that Pope Pius XII officially approved big bang theory in 1951. Most scientists were unimpressed. After all, Christianity had, with much greater fervor, asserted for centuries that the only acceptable cosmology was the Aristotelian/Ptolemaic model. Even Georges Lemaître, who was a Roman Catholic priest, took pains to separate his science from his religion, at least publicly. (On the other hand, Gamow, who enjoyed tweaking other scientists, once cited the papal approval in a technical paper.) During the first sixty years of the 20th century, there were essentially no observations that could choose among the big bang, the steady state, or other models, since all could explain the observed redshift-distance relationship, though in very dif-

The need for better data

ferent ways. The data at high redshift that could, and later did, decide between steady state and big bang were either lacking or unreliable until the 1960s. In the early days of observational cosmology, even obtaining something as seemingly straightforward as the Hubble constant was difficult; the initial work seemed to present more challenges than support for the standard models. Hubble and Humason's observations prior to the 1950s contained sufficient systematic error that a straightforward standard model would not fit; with $H_0 \sim 500 \text{ km s}^{-1} \text{ Mpc}^{-1}$, the Hubble time was simply too short to account even for the age of the Earth.

The observational situation improved slowly after the opening of the 200-inch reflector telescope atop Mount Palomar, near Los Angeles. Walter Baade showed in 1952 that a misidentification of the Cepheid variable stars in the Magellanic Clouds, upon which much of the cosmic distance ladder was ultimately based, had led to an overestimate by a factor of two in the extragalactic distance scale. In 1958, Allan Sandage established that many of the very bright stars used by Hubble and Humason as distance indicators for their farthest galaxies were, in fact, not stars at all but extended regions of very hot, ionized hydrogen gas. This result showed that the original scale was too large by at least another factor of two. These and other new results dropped the value of the Hubble constant from the original estimate to the range of 50 to 100 $\text{km s}^{-1} \text{ Mpc}^{-1}$, where it remained for the next several decades. The age problem for the standard models became less critical.

As evidence for the big bang models grew, the steady state model came under increased scrutiny. One of the positive aspects of the steady state model is that it is so eminently falsifiable. It is so tightly constrained by the perfect cosmological principle that it makes very specific predictions on many fronts. One of these is that, on average, the universe should look today much as it ever has; there should be no overall change or evolution to the cosmos. Yet evidence for just such change and evolution accumulated slowly after World War II. The new science of radio astronomy, a field which ultimately traces its ancestry to wartime radar, provided important data in the late 1950s and early 1960s. Martin Ryle claimed in 1955 that counts of radio sources as a function of redshift were incompatible with the prediction of the steady state model. The earliest data turned out to be inconclusive, but as the technique was refined, it became more and more apparent that the data were in conflict with the steady state model. Radio astronomers also discovered a new phenomenon, the quasi-stellar radio sources.² These mysterious objects emitted huge quantities of radio energy, but on optical photographic plates they appeared starlike. When Maarten Schmidt realized in 1963 that the bizarre spectra of some quasars could be explained as the familiar spectrum of hydrogen, but redshifted far more than anyone had ever imagined to be possible, it was nearly fatal to the perfect cosmological principle. The environment of the quasars, so manifestly different from that of nearby galaxies, was a clear example of change in the universe.

Evidence for cosmic evolution

²Quasi-stellar objects, or quasars, are described in Chapter 9.

The discovery of the cosmic background radiation provided evidence for the big bang

A recipe for the early universe

Quasars were abundant in the more distant past, but hardly exist at all in the recent epoch of the universe.

While this was all evidence *against* the steady state theory, it was not evidence *for* the big bang. Such observations validated the idea of a universe evolving from an initial state, but the evolution of specific structures such as galaxies and quasars could be explained in many ways without directly testing the big bang itself. Only a remnant of the very earliest era in the existence of the universe would be convincing. That evidence was the cosmic background radiation, or CBR, which was discovered in 1964 by Arno Penzias and Robert Wilson.³ This background radiation was virtually impossible to explain within the steady state model, but was a natural outcome of a hot big bang. The presence of the CBR, the lack of a physical theory to explain the creation of matter required in the steady state theory, as well as the lack of agreement with observations of galaxies and quasars, led almost all astronomers to abandon the steady state model as a possible alternative to the big bang.

By the early 1960s, cosmologists had many revolutionary ideas before them: the expansion of the universe, the origin of the elements, the implications of the early dense phase of the big bang model, and the evidence for evolution in the universe. These various lines of thought, both theoretical and observational, were merging to create a climate receptive to the discovery of the cosmic background radiation. Astronomers were just beginning to search specifically for it when it was accidentally discovered by Penzias and Wilson during their pursuit of a different project. Thus it was that cosmologists came to accept the reality of the big bang epoch of the universe, and began to study that era in earnest. The earliest stages of the universe might seem to be so unimaginably complicated that we would not even have the capability to describe them. Yet there are good reasons to believe that near its beginning, the universe was in many ways much simpler than it is today.

The further back in time we go, the smaller the scale of the universe, the greater its density, and the higher its temperature. Complex objects such as stars and galaxies had not yet formed. The universe consisted of a soup of elementary particles, interacting with one another in relatively simple ways. Because of the high density, these elementary particles constantly exchanged energy and momentum; the particles coexisted in an equilibrium defined by a single set of statistical properties. A state completely characterized by the statistical quantity we know as temperature is called **thermal equilibrium**. The simplest model we can study assumes that the early universe may be described by a state of thermal equilibrium; from the big bang until the universe cools sufficiently that the photons no longer interact with the massive particles, the temperature alone will be our guide to the many particles and their interactions. Like any model, this is subject to testing; it finds its justification mainly in the uniform blackbody spectrum of the cosmic background radiation,

³The cosmic background radiation is the subject of Chapter 14.

and in the predictions of big bang nucleosynthesis. It does introduce some problems of its own; we shall examine these failings in later chapters. But first let us see how far we can go with this simple assumption, and how well the data support it. We shall find that it will go very far indeed.

The radiation era

The presence of the CBR has important implications for the nature of the universe in its earliest times. The CBR is not a dynamically important constituent of the universe today, but this was not always the case. Although photons have no rest mass, they possess energy proportional to their frequency.⁴ As we have learned from general relativity, both energy and mass create gravity. Just as the universe contains a rest-mass density, it also contains an energy density, the latter being defined simply as the energy per unit volume. The cosmic energy density of the present universe is mostly due to the background photons. More than a billion photons are present for every particle of ordinary matter, but each photon has lost much of its energy in the overall expansion. Hence the energy density in the CBR is minuscule in comparison to the mass density of ordinary matter, and we say that the universe today is in the **matter era**, and is *matter dominated*. Early on, however, conditions were quite the opposite. For the first several thousand years of the universe's existence, the radiant energy density provided the most important contribution to the gravity; therefore, this stage in the history of the universe is called the **radiation era**. Since the rest-mass density of the matter was then entirely negligible in comparison to the energy density of the radiation, the universe is said to have been *radiation dominated*. During this era, the evolution of the cosmos was controlled by the photons.

How must the Friedmann equations be changed to account for photons? Obviously, we must include the photons' energy density in the total matter-energy density ρ ; but in this case, the pressure due to the photons is also a significant contributor to the universe. The pressure in the early universe is not a gas pressure, since contributions from the kinetic and other energies of the massive particles were largely insignificant, but a *radiation pressure* resulting from the energy density of the photons. Although they are massless, photons carry energy and momentum; thus when they impinge upon a surface, some momentum can be transferred to it, resulting in a pressure. Radiation pressure is not something that is often apparent in everyday life, but its effects can be seen in the sky; the tails of comets point away from the Sun in part because the specks of dust that make them up are buffeted by the pressure from the Sun's photons. A pressure force arises when the energy contained

In the early universe radiation was more important than matter

Radiation pressure contributes to gravity

⁴Specifically, the energy E of a photon with frequency ν is given by $E = h\nu$, where h is a fundamental constant of physics known as Planck's constant. See Chapter 4 for further details.

in a volume is able to exert a net force on some surface, which could be a surface contained within the volume and need not be a boundary. In general, a pressure can be associated with any energy density other than pure rest mass-energy. An *equation of state* is a relationship that connects pressure (force per unit area) with energy density (energy per unit volume); the ideal-gas law discussed in Chapter 5 is an example of such an equation. Cosmologists prefer to write the equation of state using the relationship $P = w\rho$, where w is the *equation of state parameter*. This form is general enough to account for any sort of energy that might fill the cosmos. For photons $w = 1/3$.

Previously, we have investigated the expansion of the universe when pressure is assumed to be negligible. We found its behavior to be very much like the prediction from ordinary Newtonian gravity in an expanding background, with some corrections for relativistic effects. A pressureless universe was easy to understand in fairly familiar terms. What happens when we introduce a pressure? One counterintuitive result of this new effect is that any ordinary positive pressure actually *increases* the gravitational force. Pressure doesn't inflate the universe, it causes deceleration! However, some reflection should make it clear that this is not so surprising after all. In general relativity, not only energy density, but also *stress*, of which pressure is one form, participates in gravity. From this the conclusion follows that pressure too must actually contribute to gravitational attraction, and thus possibly to a collapse. But does not pressure also push objects apart, and so should it not push the universe outward? Pressure only pushes when there is a *change* in pressure. Pressure pushes from regions of high pressure to regions of low pressure. In a homogeneous and isotropic universe, the pressure must be the same everywhere, so there is no net force. Pressure is left with nothing to do but increase the gravitational attraction.

The evolution of the early universe is controlled by the photons; we must thus determine the behavior of the photon energy density as a function of the scale factor. How does this energy density change with expansion? First, the wavelength of any individual photon is redshifted by the overall expansion, with $\lambda \sim R$ for any wavelength; thus the representative wavelength redshifts in exactly the same way, since R is the same for all photons at any cosmic time. Another effect to consider is that the fixed number of photons occupies a larger and larger space as the volume expands, that is, the photons become more and more diluted. Hence the *number* of photons in a volume of space decreases due to the increase in volume from the expansion. Since the volume increases as R^3 , we conclude that the number of photons per unit volume decreases like $1/R^3$, just as for matter density. Unlike the matter density, however, the photon *energy* density is also affected by the aforementioned redshift. Combining these two effects, we find that the energy density of the cosmic photons, which we designate as \mathcal{E} , decreases as the *fourth power* of the scale factor, that is,

$$\mathcal{E}(t) = \mathcal{E}_0 \left[\frac{R_0}{R(t)} \right]^4, \quad (12.1)$$

Photons in an expanding universe

where \mathcal{E}_0 is the present radiation energy density. The radiation energy density drops more rapidly than it would if its decrease were due only to the volume expansion of the universe, since the redshift introduces an additional power of the scale factor.

The energy density of radiation in thermal equilibrium is proportional to the fourth power of the temperature, $\mathcal{E} \propto T^4$. Hence the temperature of the radiation in the universe is a simple function of the scale factor

$$T(t) = T_0 \left[\frac{R_0}{R(t)} \right]. \quad (12.2)$$

Using the formula for redshift as a function of scale factor, we obtain the temperature at any redshift z in terms of the present temperature and the redshift:

$$T(z) = T_0(1 + z). \quad (12.3)$$

The CBR temperature is proportional to redshift

We need not know the time corresponding to z in order to make use of this equation. Equation (12.3) casts temperature in terms of observable quantities; T_0 is obtained from the cosmic background radiation and redshift z is measured from spectra.

As we go backward in time, corresponding to larger and larger z and smaller and smaller scale factor, we find a universe filled with photons of increasing temperature. The cosmic background radiation today has a temperature of 2.725 K above absolute zero. At a redshift $z = 1$, this same background radiation had a temperature of 5.45 K. As we look into the far distant past, the temperature rises into the thousands of degrees. The background radiation, which today is mostly in the microwave band, becomes visible; the universe once was bright with light. Further back in time, the temperature rises to ever greater values. Indeed, if the universe began with a scale factor of zero, the initial temperature must have been infinite!

The behavior for the scale factor as a function of cosmic time in the early universe is computed by solving the Friedmann equation, under the assumption that the energy density arises only from the radiation. It is easiest to solve this equation for the case of flat space. In our studies of the present universe, we have often employed for comparison purposes the Einstein–de Sitter model, the solution for flat space ($k = 0$) and no cosmological constant ($\Lambda = 0$). As it turns out, these are good approximations regardless of the actual curvature, at sufficiently early times. In the standard models, the density, from whatever source, becomes large as the scale factor shrinks; furthermore, the energy density will change with R at least as R^{-3} . Therefore, as R becomes small the energy density term, that is, the first term on the right-hand side of the Friedmann equation (11.22), will become large, whereas the curvature and cosmological constants cannot change. Eventually the density term dominates so completely that the other terms are irrelevant.

The Einstein–de Sitter model is a good approximation for the early universe

The Friedmann equation for the flat universe has the form

$$\dot{R}^2 \propto \rho R^2. \quad (12.4)$$

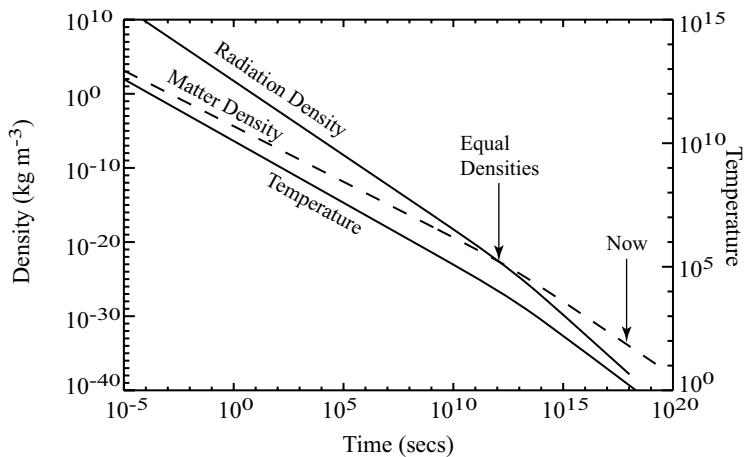


Fig. 12.3 The thermal history of the universe. The scale on the right is the temperature; the scale on the left is density. The top solid line is radiation density, the dashed line is matter density, and the bottom solid line is the temperature. The time of equal densities, when the matter and radiation densities are equal, occurs around 10^{12} s after the big bang.

The difference between a matter dominated universe and a radiation dominated universe is the behavior of ρ with R . In a matter dominated universe, $\rho \propto R^{-3}$, while in a radiation dominated universe $\rho \propto R^{-4}$. In the matter dominated Einstein-de Sitter universe, the solution to the equation has the form $R(t) \propto t^{2/3}$. We can now solve the Friedmann equation to find that during the radiation era, the scale factor is given by

$$R(t) \propto t^{1/2}. \quad (12.5)$$

The age of a flat, radiation dominated universe is only $1/2$ the Hubble time. The deceleration is larger in such a universe, with $q = 1$. As we have stated, the presence of radiation pressure actually increases the gravitational force, thus braking the expansion more rapidly.

Figure 12.3 illustrates the thermal history of the universe. Today matter dominates the universe, but as we go back to earlier and earlier times, the radiation energy density increases at a faster rate than does the matter density. Before the point of equal densities, radiation ruled the universe.

Matter and energy

What would such a radiation dominated universe be like? High temperatures and energies imply drastically different conditions from what we know today. The further we probe toward $t = 0$, the more exotic the universe becomes. At sufficiently early times, conditions were so extreme that even atoms could not have existed. The universe was like a tremendous particle accelerator, with high-energy particles zipping about at relativistic speeds, crashing into one another and interacting with photons. Our description of the very earliest moments of the universe must necessarily be somewhat tentative, as the theories of matter and energy under such extreme conditions are still rudimentary. But let us see how far we can go with the knowledge we have.

Particles, antiparticles, and energy

If the very early universe was filled with particles, from where did they originate? As has been amply demonstrated experimentally, Einstein's famous equation $E = mc^2$ means that matter can be converted into energy. What may not be so well appreciated is that it also goes the other way: pure energy can be converted into matter. We have seen an example of this phenomenon in our study of Hawking radiation from a black hole, but other such processes exist. In the early universe, creation of matter from energy was one of the most important effects. The rest mass of any elementary particle is equivalent to some amount of energy, defined to be the rest energy of that particle. In the state of thermal equilibrium, the total energy is divided equally among all species of particle, including photons. At any cosmic time in the early universe, the temperature implies a mass scale, via this mean energy per particle. If the temperature of the early universe is at or above this *threshold* value for a given particle, two colliding photons can produce the particle and its antiparticle, a phenomenon known as **pair production**. The threshold temperature thus represents the minimum energy required for matter–antimatter partners of a specific mass to be created from the collision of two photons. Particles of a given species can still be produced at temperatures well above their threshold, of course; in this case, they are simply created with kinetic as well as rest energy. Pair production need not always result from photon collisions; if the temperature is at least twice the threshold, particle pairs can appear directly from the energy of the electromagnetic fields.

In thermal equilibrium, the mean energy per particle is proportional to the temperature

$$\langle E \rangle = \frac{3}{2}k_B T. \quad (12.6)$$

As in the ideal-gas law, Boltzmann's constant k_B appears as a conversion factor between temperature and energy units. To find the threshold temperature corresponding to any particle rest mass m_0 , equate the rest energy of the particle with the mean energy of the photons and solve for the temperature:

$$T = \frac{2m_0c^2}{3k_B}. \quad (12.7)$$

There is a simple direct proportionality between temperature and particle mass. The higher the temperature, the more massive the particles that can be produced. As an example, if we wish to create a proton–antiproton pair by photon collision, the temperature must be approximately 10^{13} K.

At temperatures above the threshold for a particular particle type, all reactions proceed at rates precisely equal to their backreactions; creation and destruction of any kind of particle must be exactly in balance. In pair production, for example, photons produce pairs, and the pairs annihilate back into photons. This leads to a state in which the number of particles is always nearly equal to the number of photons. Any excess of photons would create more particles, causing the number of particles to rise, whereas a shortage of photons would mean that particle creation

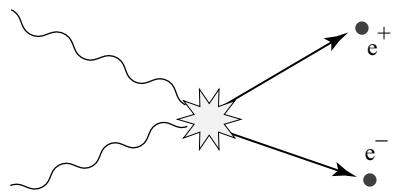


Fig. 12.4 An example of the process of pair production. Two high-energy photons collide and produce an electron–positron pair. Each photon must have at least as much energy as the rest energy of the electron, according to Einstein's formula $E = mc^2$.

Threshold temperature for pair production

would not be able to keep up with particle destruction. At any cosmic time in the early universe, then, all particles permitted to exist at the corresponding temperature were constantly colliding, materializing, and annihilating, such that the number of members of any particle species remained the same, and each particle was matched by an antiparticle. During epochs of particle creation, the contribution of matter to the energy density of the universe was thus not negligible. When the temperature dropped below the threshold temperature for a particular pair type, the annihilation rate exceeded the production rate, and the pairs of particles were rapidly destroyed.

In the ordinary matter creation from photons which we have described, both a particle and its antiparticle *must* be created, because in any such reaction, certain properties of the particles, as well as the total mass-energy, must remain the same. Nevertheless, matter and antimatter cannot have been in exact balance for all times, because the universe now is filled with matter even though the radiation temperature has dropped to 2.725 K, far below the threshold temperature for any of the constituent particles of atoms. If nucleons and antinucleons had occurred in equal numbers in the early universe, they would have annihilated, leaving only photons. Considerations of causality indicate that matter and antimatter could not have segregated themselves fast enough in the early universe to prevent this. Furthermore, there is no evidence for any significant accumulations of antimatter anywhere in the present universe. Galaxies and clusters do not exist in isolation; if an antimatter galaxy existed, it would undoubtedly find matter in short order. Any such antimatter would annihilate whenever it encountered matter, creating a copious flux of characteristic gamma rays. Matter–antimatter reactions are particularly violent, and can be seen over enormous distances if they occur. No tremendous cosmic flux of gamma rays from annihilation reactions is seen, and we can conclude that only matter exists in appreciable amounts.

Since our present universe is made of matter, at some point in the early high-temperature epoch, the perfect balance between matter and antimatter must have been violated. The amount of leftover matter is rather small; there is one particle of matter per 1.6 billion photons, meaning that the excess of matter over antimatter was about one part in a billion. Yet that small quantity of matter makes up all that we can see, and all that we are. There is as yet no firm explanation of how this effect occurred. While it could have been simply an initial condition, built into the very beginning itself, physicists believe that it may be a consequence of the fundamental laws of physics; one leading possibility will be discussed when we delve into the chronology of the big bang.

The universe today contains photons and leftover matter

Fields of dreams

The extreme conditions of the early universe require that our understanding of cosmic history must be inextricably linked with particle

physics. Throughout the 20th century, there were occasional interactions between cosmology and other branches of physics; some of the most distinguished physicists of the first half of that century, such as Enrico Fermi, George Gamow, Robert Oppenheimer, and many others, made important contributions to cosmology. But for the most part, nuclear and particle physics advanced independently of cosmology. Particle physicists have sought to build ever larger accelerators, in order to study physics at higher energies. But such accelerators take an increasing toll in effort and resources. Our study of special relativity showed how difficult it is to accelerate even elementary particles to relativistic speeds; if we wish to create even more exotic states, such as significant quantities of antimatter, the engineering problems become considerable, even overwhelming. The largest accelerator ever planned, the Superconducting Supercollider, was to have consisted of an evacuated ring, 54 miles in circumference, about which nearly infinitesimal particles would have been driven to ever higher energies by the magnetic field from superconducting magnets. The cost of this great machine proved prohibitive, however. And even the Supercollider could not have reached the energies for which the particle physicists ultimately yearn. In order to test the leading edge of particle physics to the utmost, much greater energies are necessary. With any realistically foreseeable technology, only the early universe itself could be an appropriate laboratory.

Describing the early universe requires a theory of high-energy physics

The world of elementary particles is a realm controlled by **quantum mechanics**, the physics of the very small. In quantum mechanics, the sureties of our familiar, macroscopic world vanish, to be replaced by a physics in which only probabilities can be known. We cannot predict, for example, when a given atom of uranium will decay; not because of any lack of understanding about the decay process, or ignorance of the initial state of the atom, but because it is *fundamentally* unknowable. The best we can do is to compute the probability that the atom will decay in any specified interval of time. If we have a large number, that is, an *ensemble*, of uranium atoms, then we can predict how many will have decayed after a particular time interval has elapsed, but we can say nothing definite about the fate of any individual atom.

The strange world of quantum mechanics

Quantum mechanics also demands a blurring of the concepts of “particle” and “wave.” According to quantum mechanics, each entity can show both corpuscular and wave behaviors, though never both at once. We are already familiar with something that can show either wave or corpuscular behavior: light. The typical wavelength of a photon of visible light is about 5×10^{-7} m, a length that is greater than the size of molecules. This means that visible light often manifests itself as a wave as it interacts with surrounding matter. Yet we have often explicitly treated light as a particle, the photon, such as when we deal with quantum transitions in atoms. We are less familiar with the wave nature of things we call particles because their wavelengths are so small. In general, the wavelength of a massive particle depends upon its velocity, but high-energy particles moving at relativistic speeds manifest a

Wave-particle duality

wavelength known as the **Compton wavelength**,

$$\lambda_C = \frac{h}{m_0 c}. \quad (12.8)$$

For example, the Compton wavelength of a proton is approximately 2×10^{-16} m. Although our mental picture of entities such as the proton and the electron is firmly rooted in the concept of particles, their wave nature is easily observed in high-energy physics experiments. Wave-particle duality is as real for the proton as it is for the photon.⁵

In modern theories of quantum mechanics and particle physics, a wave, and hence its allied particle, can be associated with a **field**. In physics, a field is a convenient mathematical representation of a quantity that is extended in space and/or time. The gravitational and electromagnetic fields are familiar descriptions of the corresponding forces. What may not be so obvious is that these fields are associated with particles. The photon provides a somewhat concrete illustration; the photon, the particle of light, is linked with the electromagnetic field. In *quantum field theory*, this concept is further extended. Every particle has an associated field; there is an electron field and a neutrino field and so on. A few of these wave-particle fields *mediate* forces.⁶ The field of an as-yet undiscovered particle, the **graviton**, is the gravitational field. The energy and momentum of a field is *quantized* into bundles, or *quanta*; the quanta manifest themselves as the particles corresponding to the field. In this view, photons are simply the quanta of the electromagnetic field itself.

The density of quanta determines the strength of the field. When this density is sufficiently high, the quantum nature of the field is disguised and classical field theory is valid. For example, the surface of the Sun spews forth so many photons each second that the electromagnetic field (the light) emanating from it seems continuous and thus obeys the laws of classical optics. Yet many photoelectric instruments, such as the charge-coupled devices (CCD's) used in modern telescope detectors and digital cameras, function by interacting with discrete photons; sufficiently sensitive devices can detect electromagnetic fields so weak as to represent only a few photons impinging upon the detector. The rod cells in the retina of the vertebrate eye achieve such a level of sensitivity; nature discovered the principle long before humans incorporated it into modern technology. The retina is lined with cells containing special molecules that can, upon being struck by photons, change their shapes. The alteration in the configuration of such a molecule rearranges its electrical charges, and thus creates a weak electric current. Ultimately, after considerable amplification and processing by nerves, the brain interprets such currents as an image. The retina is a device for converting a quantum field into a pattern of electrical activity which a processor, the brain, can recognize! Yet when the light enters the eye, passing

⁵Wave-particle duality is discussed further in Chapter 17.

⁶Carrier particles and their role in mediating forces are discussed briefly in Chapter 4.

through the cornea and the lens, it behaves like a wave, and we can use classical wave optics to compute how much it will be refracted by the cornea and lens and where the focus will be; if necessary, we can then interpose an artificial lens between the source and the eye in order to shift the point of focus to the retina. In one part of the eye, light behaves like a wave; in another, the very same light is a particle. According to quantum mechanics, it is either, depending upon which behavior the experiment elicits. Quantum mechanical effects are not weird theoretical constructs with no connection to reality; this *is* the way the universe works. Quantum mechanics, perhaps even more so than relativity, is very nonintuitive. But whether we are aware of it or not, it impinges in many ways upon our classical, macroscopic world.

Field theories of one form or another are the foundation of most of modern particle physics. Some of the most important characteristics of the fields are their **symmetries**, those quantities that remain invariant under specific transformations. We have already talked about invariance in relativity; there we can find the most intuitive forms of symmetry. The space-time interval is unchanged when the coordinates change. The cosmological principle is a statement of the symmetry of the universe in both spatial location (homogeneity) and spatial direction (isotropy). Many conservation laws can be attributed to various symmetries. For example, energy conservation arises from symmetry with respect to translations or reversals in time. Energy can be defined to be a particular quantity that does not change as time changes; the fact that such a quantity can be specified at all is due to the existence of the symmetry in time of fundamental laws of physics. Similarly, the conservation of linear momentum can be understood as resulting from symmetry under straight-line translations in space, while conservation of angular momentum is a consequence of symmetry with respect to rotations in space.

Particle physics itself can be characterized as a search for symmetry. The menagerie of particles can be divided into a few families, each with various *internal* symmetries, that is, symmetries which are properties of the field itself and not of the outside world. Electric charge, for example, represents a symmetry in the electromagnetic field under certain transformations of abstract coordinates. Three of the four fundamental forces of nature can be understood in terms of the symmetries of an appropriate field theory. (Gravity is the lone holdout, so far resisting all efforts to fit it into this picture.) If these forces share all symmetries, then they are indistinguishable from one other. This was the case in the earliest times of the big bang. Today we see the forces as distinct because of the loss of symmetries at early times, once the temperature dropped below certain levels. Such a loss of symmetry is known as **spontaneous symmetry breaking**.

It may be difficult to visualize a spontaneous symmetry breaking of some abstract field theory, but we are all familiar with a very similar symmetry breaking: freezing. A liquid, such as water, has higher temperature and greater disorder. It can be rotated or translated by any

Conservation laws reflect underlying symmetries

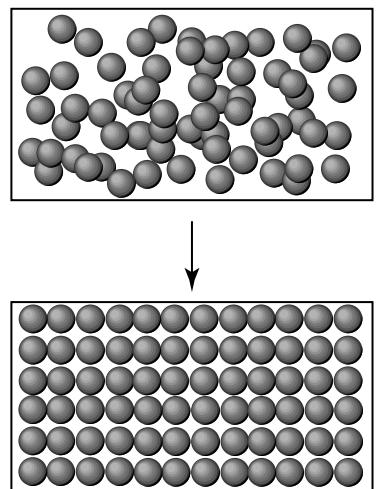


Fig. 12.5 In freezing, molecules go from a disordered, completely symmetric state to an orderly arrangement of very specific symmetries. Thus, the phase transition of freezing results in a loss of symmetry.

amount and it will still look the same. When the temperature is lowered and the water freezes, it forms a crystal lattice that has much greater structure; it is symmetric only under rotations of particular angles, or translations equal to the lattice spacing. Similarly, under appropriate conditions, forces are united in a symmetry; but failure of those conditions, such as a lowering of temperature in the case of fundamental forces, breaks the symmetry. The freezing of water is called in physics a *phase transition*; boiling of water from a liquid to a gas is also a phase transition. Remarkably, it is not only an aid to the imagination to visualize spontaneous symmetry breaking as similar to such mundane occurrences; the two concepts are actually *mathematically* quite similar. It is thus quite justifiable to think of a spontaneous symmetry breaking as a kind of phase transition.

The search for a “Theory of Everything”

The ultimate goal, the Holy Grail of particle physics, is the “Theory of Everything,” a theory that would encompass all particles and forces, showing them to be manifestations of an underlying simplicity. Like the Holy Grail of mythology, the final theory has proved elusive. Many have thought to grasp it, only for their vision to evaporate in the glare of data. It may be that a few have glimpsed its outline, but as yet no one has seen it clearly. But the quest continues, and it may someday be successful, for much has already been learned. As new and more powerful particle accelerators were constructed, particle physics contributed new ideas and discoveries to cosmology, until the universe seemed comprehensible down to the first hundredth of a second. Now it may be that cosmology can return the favor, by providing clues toward the understanding of conditions that may never be reproducible by humans.

The beginning of time

We would like our cosmological theory to describe the history of the universe all the way back to the big bang, with its soaring temperatures. But can classical general relativity apply as far as $t = 0$, with its extraordinary conditions? Certainly it cannot. Gravity has not yet been fully incorporated into the other great theory of modern physics, quantum mechanics. For the conditions prevailing throughout most of the history of the universe, we can separate the two theories under almost all circumstances, because the scales over which they dominate are so vastly different. Quantum mechanics rules the smallest scales, while gravity governs the largest scales. In the present universe, there is little overlap in the domains of these two theories. But as we approach $t = 0$, their regimes must merge together.

Quantum gravity would describe the earliest moments we can imagine

Gravity controls the evolution of the universe because it is long range, and especially because it is *only* attractive; all the mass in the universe contributes, and the gravitational force is never partially cancelled by any negative gravitational charge. Any repulsive counteraction to gravity must come from a cosmological constant. Yet gravity is by far the weakest of the four fundamental forces. Even at quite early times in

the history of the universe, gravity was much weaker than any of the other fundamental forces, so we may go to very early epochs before we must be concerned about quantum gravity. At the very beginning of the universe, however, the scales characteristic of quantum mechanics and of gravity were similar, and gravity was comparable in its immediate effects to the other forces. We must have a full theory of quantum gravity in order to describe the universe under such conditions. Since we have no such theory, the earliest moments of the big bang remain a mystery.

Rather than starting from time zero, then, we must pick up the story where classical general relativity gains control of the universe as a whole. This occurs at a cosmic time of 10^{-43} s, the *Planck time*. At this time, the characteristic length-scale of the universe was $ct = 1.6 \times 10^{-35}$ m, the *Planck length*.⁷ This length is much smaller than the Compton wavelength of any elementary particle. Indeed, the very idea of a particle, at least as we currently conceive of it, must break down during this initial period. We can say essentially nothing about the behavior of the contents of the universe from the beginning until the Planck time, an interval which is often called the **Planck epoch**. During this epoch, all four fundamental forces of nature (gravity, electromagnetism, and the weak and strong nuclear interactions) composed a single force. At the end of the Planck epoch, the gravitons fell out of equilibrium with the other particles, and gravity decoupled from the other forces. The gravitons then streamed out through the universe, forming a cosmic background of gravitational waves. This event occurred so early that the energy of these waves has been redshifted nearly away, and they are utterly undetectable today; the invaluable information that they could provide about the earliest moments is beyond our grasp for the foreseeable future. The decoupling of gravity was the first spontaneous symmetry breaking in the universe, the loss of the perfect symmetry and equivalence among all four forces with which the universe is thought to have begun. As strange as it may seem, at its very beginning the universe, with its exotic conditions we cannot yet comprehend, was in some ways as simple as it could ever be.

From the Planck time till about 10^{-35} s, the temperature was so high that we still have little understanding of the nature of matter under these conditions. This interval can be called the **unified epoch**, since during this stage electromagnetism, the weak interaction, and the strong interaction were unified; that is, they made up a single, indistinguishable force. Although theories exist that apply to conditions during the unified epoch, they are still incomplete and are not always consistent with experimental data. Nevertheless, they provide the beginnings of a framework to understand the behavior of particles and forces during this epoch. These theories are called **grand unified theories** (GUTs) because they attempt to explain the unification of the three forces. Unfortunately, the temperatures, and hence the energies, characteristic of

The unified epoch

⁷These scales are named for Max Planck, in honor of his early contributions to quantum mechanics.

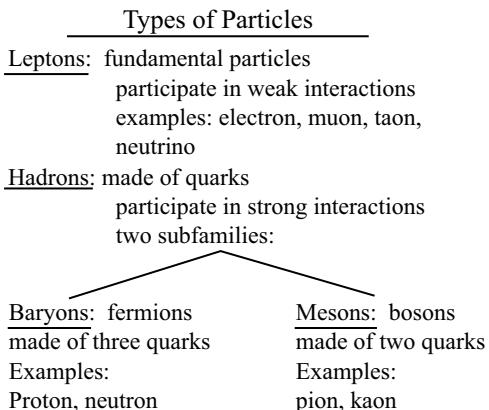


Fig. 12.6 Types of particles. The hadrons participate in the strong interaction; the leptons do not. Hadrons are composed of quarks and are subdivided into the baryons and the mesons. The proton and neutron, the constituent particles of atomic nuclei, are the most important examples of baryons.

The time of inflation

this epoch are far beyond what we could ever hope to reach in the largest particle accelerator we could imagine. Physicists and cosmologists hope that the universe itself will provide experimental evidence for conditions during the unified epoch, and thereby guide the development of GUTs.

Sometime before the end of the unified epoch, the universe underwent a startling transition. Cosmologists believe that the universe entered a period of exponential expansion called **inflation**. The universe was filled with an energy associated with empty space, a form of a cosmological constant. A model with properties such as these has been previously encountered in the form of the de Sitter universe.⁸ If inflation occurred, it must have taken place sometime around 10^{-37} s after the big bang. This exponential expansion lasted for only a brief time, but the interval was long enough to inflate the universe by an enormous factor of e^N , where N is a number at least as large as 100 and perhaps as great as 1000. At the end of this inflationary period, the energy in the vacuum of space was converted into more conventional forms of matter and energy. This event is known as *reheating*; it marks the beginning of standard cosmological evolution.⁹

The most significant remnant of the unified epoch is the excess matter remaining after the epoch ends. The elementary fermions that make up ordinary matter in the present universe can be subdivided into two families, the **hadrons** and the **leptons**, according to whether they respond to the strong interaction. Hadrons participate in strong interactions, while leptons are governed by the weak interaction. Hadrons are not themselves point particles, but consist of smaller particles called **quarks**, which do seem to be pointlike. Six species of quark were predicted from theory; five were found fairly easily in high-energy particle physics laboratories, with reliable evidence for the sixth beginning to appear early in 1995. Based upon their construction, and the species of quarks present, the hadrons may be further broken down into the **baryons**, which con-

⁸See Chapter 11.

⁹Inflation will be discussed in more detail in Chapter 16.

Elementary particles

sist of three quarks, and *mesons*, which are composed of a quark and an antiquark. Mesons are heavy particles with extremely short half-lives; we shall have little further to say about them. Baryons, on the other hand, are extremely important; by far the most common baryons are the nucleons, the proton, and the neutron. The term baryon is sometimes even loosely used as a synonym for nucleon, although this is not quite correct. Many other baryons exist, but in the present universe they are rare and decay very quickly, eventually always becoming protons, since the proton is the least massive baryon. During the unified epoch, however, conditions were so extreme that quarks had not condensed into hadrons; the universe consisted of a brew of highly relativistic particles, including quarks and more exotic particles.

Particles created from pure energy in ordinary processes must always be created in matter–antimatter pairs. Furthermore, when a particle and its antiparticle collide, they destroy one another, converting their rest masses entirely into photon energy. This corresponds to a rule called *conservation of baryon number*, where antiparticles of baryons are negative baryons. For example, if a single neutron, a baryon, decays, only one proton can be created. If there were no baryons to begin with, then only particles that are not baryons, and thus have zero baryon number (such as photons), or else a pair consisting of both a baryon (baryon number +1) and its antiparticle (baryon number –1) must be created. Since baryons are made of quarks, this also implies conservation of the number of quarks in a particle reaction, and under ordinary conditions this is true. However, a common feature of most GUTs is that this particular conservation law no longer holds; reactions can take place that transform quarks into leptons and *vice versa*, thus violating baryon conservation. These reactions occur in such a way that the result is always a tiny excess of matter. The process by which matter was preferred over antimatter, creating the stuff of our universe, is called **baryogenesis**.

If baryogenesis had *not* occurred, whether by the GUT mechanisms or by some other means, no matter would now exist, for every particle would have eventually destroyed itself with its antimatter partner. In the present universe, for every hadron there are a little over a billion (more exactly, approximately 1.6×10^9) photons left over from the early seconds of the big bang; therefore, this asymmetry between matter and antimatter during the unified epoch must have been at a level of only about one part in a billion. Yet it is just this asymmetry that led to the creation of all the many forms of matter that we know today.

Baryogenesis in GUTs leads to another very important prediction: the proton is unstable. There is no particular reason that the proton *must* be stable. After all, the proton is not a truly “elementary” particle, but is a composite of three quarks. However, protons are one of the most important components of ordinary matter, so the question of their stability is of considerable significance for the stability of matter as a whole. We can immediately see that the proton must have an extremely long life expectancy, for if it did not, matter would disintegrate over the current age of the universe. If the life expectancy of the proton were

A slight excess of matter over antimatter created the material contents of the universe

short, comparable to the age of the universe, then given the vast number of protons within our Hubble sphere, we would expect to be able to see proton decay on a regular basis. In particular, the human body contains approximately 2×10^{28} protons. If the proton's life expectancy were comparable to the age of the universe, approximately 10^{10} years, then on average, roughly 10^{18} protons would decay in an individual body per year! Since the decay products would have considerable energy and would rapidly be converted into gamma rays, ordinary matter would be noticeably radioactive. Life as we know it probably could not exist in a universe with such a short proton lifetime. On the other hand, a proton lifetime that is considerably greater than the age of the universe would mean that very few protons would have decayed by now, explaining the absence of an observable effect.

As it happens, testing this prediction is an experiment that does not require a particle accelerator at all; it is one of the few aspects of GUT theories that can be directly tested in Earthly laboratories. It is easy to design an experiment to measure the life expectancy of the proton. Simply gather a large number of protons (a large quantity of water will work nicely) and watch for decay products. For example, if the specimen has approximately 10^{30} protons, and the proton's half-life is 10^{30} years, then roughly one proton decay per year should be observed. Experiments of this nature have been performed, and the average lifetime has proved to be too large to measure with current technology. A firm lower bound can be placed: the proton lives at least 10^{31} years, probably closer to 10^{32} years; it might even be stable. This result disagrees with the simplest GUT theory, which predicts a proton lifetime of about 10^{30} years. This does not rule out all GUTs, but it means that the simplest version cannot be correct. More sophisticated, and thus more difficult, theories are required. Even so, simple GUTs have some successes, such as explaining why matter exists. We may not yet fully understand the unified epoch, but there is good reason for optimism that it will become comprehensible in the near future.

Quarks, hadrons, and leptons

The end of the unified epoch came at 10^{-35} s, when the temperature dropped below the level required to maintain the grand unified symmetry, and the strong interaction decoupled from the other forces. What followed may be called the **quark epoch**. In the present universe, solitary quarks do not exist in nature; particles called **gluons** hold quarks together in hadrons. During the quark epoch, however, the universe consisted of free quarks and gluons, along with the carrier particles of the combined electromagnetic and weak force, as well as more exotic heavy particles; plus, of course, the antiparticles of all. We still cannot say very much about this period. Although a theory of strong interactions, *quantum chromodynamics*, or QCD, exists, its equations are so complicated that very little is known about their solutions; this difficulty

occurs precisely because the coupling between two hadrons is so strong. We know even without any equations that the strong interaction is exceedingly strong; after all, it holds together nuclei against the protons' electrostatic repulsion. But the information so far extracted from the theory has revealed a most curious property of this force: it actually becomes stronger with increasing distance. This is why free quarks are never found under natural conditions in the present universe; if any two quarks were somehow separated, the force between them would increase until the energy in the strong field would create a new pair of quarks. It would be akin to trying to divide a magnet; when a bar magnet is split, the result is two smaller magnets, not two distinct poles. At extremely high energies and densities, however, the strong nuclear interaction becomes negligible, and the quarks are able to behave as if they were perfectly free particles. Thus under the conditions of the quark epoch, there was no compulsion for the quarks to form particles.

During most of the quark epoch, the weak and electromagnetic forces were unified as the **electroweak interaction**.¹⁰ At sufficiently high temperatures, the weak and electromagnetic forces were of comparable strength. Rather than photons, two other force-carrying particles, both massless, were present. Around a time of 10^{-11} s and a temperature of 10^{15} K, the weak interaction decoupled from the electromagnetic force, leaving all forces separated as they are today. During this transition, the carrier particles of the unified electroweak force were transformed into four new particles: the W^+ , W^- , and Z^0 , which acquired mass, and the photon, which did not. The three massive particles are carriers of the weak interaction, whereas the familiar photon is the carrier particle of the electromagnetic force. Because of this, the range of the electromagnetic force is, in principle, infinite, whereas the weak interaction has a short range. The masses of the W and Z particles are comparable to the masses of fairly heavy atomic nuclei. A large mass corresponds to a very high energy; hence it was difficult to create W and Z particles even in advanced accelerators. Indirect tests confirmed the electroweak theory, but the final proof had to await detection of the particles themselves. The W and Z particles were found in experiments at the CERN accelerator near Geneva, Switzerland in 1983, more than 15 years after the theory was first proposed.

The weak and electromagnetic forces were unified during the quark epoch

The symmetry breaking of the electroweak force was, like other spontaneous symmetry breakings in field theories, analogous to a phase transition such as the freezing of ice. One characteristic of phase transitions is that local conditions may affect quantities such as the rate or timing of the transition. Nearly everyone has seen a frozen creek or lake whose ice contains dark planes separating regions that froze at various rates or times, or even including pockets of liquid coexisting with the solid state. Similar phenomena could occur in spontaneous symmetry-breaking transitions; the universe may be divided into many domains

¹⁰The 1979 Nobel Prize in physics went to Steven Weinberg, Abdus Salam, and Sheldon Glashow for their work in developing the electroweak theory.

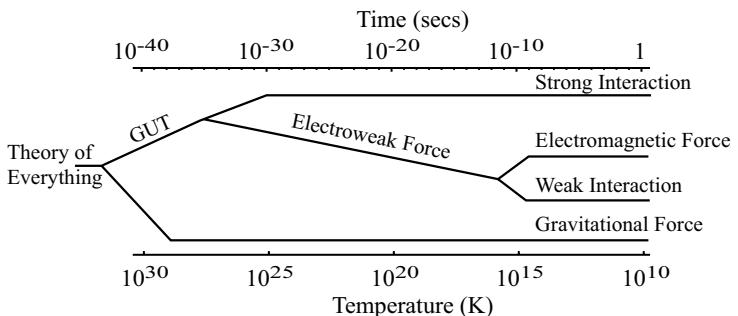


Fig. 12.7 Force tree showing how the four fundamental forces in existence today separated as time advanced in the early universe.

in which the electroweak transition occurred differently as the universe cooled. Such divisions between regions would be *defects* in the structure of space-time, similar to the defects in a crystal that divide one ordered area from another. If these defects exist, they could have attracted matter, providing seeds for later gravitational collapse.

After the separation of the fundamental forces, matter continued to exist in the form of the quark-gluon plasma. Reproducing this state could give valuable clues to the nature of matter; several heavy-ion accelerators around the world, especially at CERN and at Brookhaven National Laboratory, have been attempting to create a quark-gluon plasma by colliding heavy nuclei, such as gold or lead, at relativistic speeds. No unequivocal evidence for the creation of this substance has yet been recorded, but there is considerable optimism that it will soon be established. The quark-gluon plasma was the precursor to the next major stage in the evolution of the early universe. Approximately 10^{-6} s after the big bang, the quarks condensed into hadrons, ushering in the **hadron epoch**. The hadron epoch was brief, for the temperature soon fell below the threshold for protons. Once this occurred, the asymmetry of matter remaining from the GUT epoch was permanently frozen; all baryon-antibaryon pairs disappeared, leaving behind photons, while unpaired baryons survived. Those photons now make up most of the cosmic background radiation. Today, there are just over a billion and a half cosmic photons per baryon of ordinary matter. But as cosmic time has passed, the photons have lost their energy to the redshift. The baryons, on the other hand, retained their rest mass unchanged. This leftover bit of matter, the one-part-in-a-billion survivors at the end of the hadron epoch, went on to dominate the universe until the present.

After the condensation and decoupling of nucleons, at around 10^{-4} s, the universe entered the **lepton epoch**, when particles associated with the weak interaction ruled the cosmos. Leptons are the lighter fermions, the electrons, muons, and neutrinos.¹¹ Electrons have negative charge and are stable, muons are more massive, have a negative charge, and are unstable, while neutrinos are, as their name indicates, neutral. Each of these particles has an antiparticle of opposite charge. Like quarks, lep-

The hadron epoch

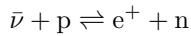
The lepton epoch

¹¹Lepton comes from the Greek *leptos*, meaning thin or small.

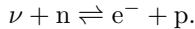
tons appear to be true point particles, genuinely elementary. In contrast to hadrons, leptons do not participate in the strong interaction, but do take part in the weak interaction.

The electrical neutrality of neutrinos renders them immune to the electromagnetic force, in addition to their unresponsiveness to the strong interaction. Neutrinos were long assumed to be massless, but there is now evidence that they have nonzero, albeit tiny, rest masses. The feebleness of the interactions of neutrinos with ordinary matter makes them exceedingly difficult to detect. At each moment, a fantastic number of neutrinos courses through a human body, originating from the Sun and from the general cosmic background. Given the average number of protons in a human body, the expected rate of its interaction with a neutrino from the Sun is approximately once every 72 years. The rest of the time, neutrinos constantly pass through us, like visible light through a pane of glass.

As cosmic time advanced and the temperature continued to fall, particles of smaller and smaller mass dominated. Early in the lepton epoch, the heavy leptons, such as the muon, were created. As the temperature dropped, production of muons essentially ceased, but muon–antimuon annihilations continued, so electrons became dominant. The universe consisted primarily of a soup of photons, neutrinos, electrons, and positrons, with the relatively small density of leftover protons and neutrons from the previous unified and hadron epochs. These sparse hadrons interacted with the leptons according to such reactions as



and



The first equation states that an antineutrino (ν is the symbol for a neutrino, and the overbar indicates the antiparticle) reacts with a proton (p) to create a positron (e^+) and a neutron (n); or, conversely, the positron and neutron can react to create a proton and an antineutrino. (The double-pointed arrow indicates that the reaction can proceed in either direction.) Similarly, the second equation says that a neutrino and a neutron can react to form an electron and a proton, or *vice versa*. These reactions must occur as indicated, because of the requirement to conserve certain properties of the particles, such as electric charge and baryon number; that is, the neutron cannot react with the antineutrino to form an electron. At high temperatures these two reactions together produced approximately equal numbers of protons and neutrons, but as the temperature continued to fall, around $t = 0.1$ s, the small difference in the masses of the proton and neutron began to have an effect. Because the neutron is somewhat more massive than the proton, at temperatures well below the threshold of either particle the reaction that produces a proton from a neutron is slightly more energetically favorable than is its counterpart that produces a neutron from a proton. Hence the interactions of the nucleons (the protons and neutrons) with the leptons

led to a much larger number of protons than neutrons, even though the hadron epoch ended with essentially equal numbers of both. The ratio of neutrons to protons continued to drop until the end of the lepton epoch.

Neutrinos decouple from the radiation

Approximately one second after the big bang, when the temperature had fallen to 10^{10} K, the density dropped enough that neutrinos no longer interacted sufficiently with other particles to remain in thermal equilibrium, and the neutrinos streamed freely from the background stew. These neutrinos continue to travel through the universe today, much like the photons of the cosmic background radiation. Unlike the CBR, however, a cosmic neutrino background would be impossible to detect directly with current technology. Neutrinos interact too weakly with other forms of matter; they are very difficult to see even when they have high energies, and the lower their energy, the less they interact. Nevertheless, because these cosmic neutrinos are very abundant, they could play a substantial role in the evolution of the universe even though their individual masses are very small.¹²

When the temperature fell below the threshold temperature of roughly 5×10^9 K for the creation of electrons, at $t \simeq 14$ s, the lepton epoch ended, fixing the ratio of protons to neutrons. This ratio is measured today to be approximately 14% neutrons to 86% protons and, at this point in the history of the universe, it had significant consequences for the subsequent formation of atoms. Almost all the leptons annihilated, leaving only enough electrons to balance the protons. The last burst of electron–positron annihilation added energy to the photons, raising their temperature somewhat, but not affecting the neutrinos, which had previously gone their own way. Because of this, the temperature of the photons at the end of the lepton epoch was 40% higher than the temperature of the neutrinos.

Nucleosynthesis

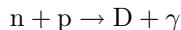
The nucleosynthesis epoch

Approximately 180 seconds after the big bang, the temperature of the universe was $\sim 10^9$ K. The contents of the universe consisted of free-streaming neutrinos, photons, and a relatively small abundance of massive particles that were mostly still in thermal equilibrium with the photons. The temperatures and densities were very high, but had dropped sufficiently that the nuclei of atoms could remain stable. *Nucleosynthesis*, the creation of atomic nuclei through nuclear reactions, commenced at this point; hence this period in the big bang is known as the **nucleosynthesis epoch**.

At high temperatures and densities, neutrons and protons can fuse directly to form **deuterium** nuclei, or *deuterons*. Deuterium, also called heavy hydrogen, is the isotope of hydrogen that contains one proton and one neutron in its nucleus. The number of protons determines which

¹²Chapter 15 will discuss this in more detail.

chemical element a given nucleus represents, but the number of neutrons affects the nuclear properties of the isotope, such as the nuclear reactions, if any, in which it will participate. Deuterium is formed by the reaction



Deuterium creation

where γ represents a photon; this reaction liberates the *binding energy* of the deuterium nucleus in the form of a photon. Under the conditions prevailing early in the nucleosynthesis epoch, deuterium readily fuses with a proton, or with another deuteron, to form the helium nucleus ^3He , or else with a neutron to form tritium, ^3H . Both of these nuclei can then react with additional particles, the ^3He with a neutron or a deuteron, and the tritium with a proton or a deuteron, to form ^4He , the most common isotope of helium. Almost all the helium in the universe, including that in the Sun, was created in this epoch, shortly after the big bang.¹³

Before this point in the history of the universe, any deuterons that might have formed were blasted apart almost immediately by the high-energy background photons; they had no opportunity to participate in any further nuclear reactions. Once the universe had cooled to approximately 10^9 K, however, some deuterons could survive. At the end of the hadron epoch, the numbers of neutrons and protons were approximately in balance. If there had been no hot photons, then all the protons in the universe would have combined immediately with the neutrons, subsequently continuing to fuse on to helium and leaving no hydrogen behind. It was this realization that led Gamow, Alpher, and Herman to propose that the early universe must have contained a billion photons per particle of matter. They also recognized that after nucleosynthesis ceased, these photons would continue to permeate the universe, redshifting to ever-lower temperatures. From this, they predicted the existence of the background radiation, more than a decade and a half before the CBR was actually discovered and fully two decades before the detailed thermal history of the early universe began to be understood in terms of particle physics.

Before nucleosynthesis began, matter consisted of a mixture of free neutrons, protons, and other particles. The amount of helium created during the nucleosynthesis epoch is not very sensitive to the density of the matter, but depends mainly on the ratio of neutrons to protons at the beginning of this epoch. However, a free neutron is unstable; it decays into a proton and an electron.¹⁴ While nucleosynthesis was progressing, two processes involving neutrons were occurring: neutrons fused with protons or deuterons, while free neutrons decayed. The competition between these two phenomena controlled the eventual abundance of he-

The creation of helium

¹³The helium found on Earth, on the other hand, is nearly all due to the radioactive decay of atoms deep in the Earth's interior; our primordial helium is long gone. The Earth is too warm, and its surface gravity is too weak, to retain atmospheric helium.

¹⁴Under laboratory conditions, this occurs with a half-life of about 10.5 minutes; that is, after 10.5 minutes, half the neutrons in a sample will have been converted into protons and electrons.

lum. Since neutrons were already relatively rare compared to protons, and since any neutrons that did not fuse with protons decayed, single protons were left as the most abundant nucleus. (A lone proton is a hydrogen nucleus.) Even so, about 25% of all baryonic mass ended up in the form of helium by the end of the nucleosynthesis epoch, which represents a significant amount of nuclear fusion. The fusing of so much hydrogen into helium did increase the temperature of the universe somewhat, but the temperature was already so high that the energy released in the fusion reactions had only a very small effect.

Hydrogen is now by far the most common element in the universe, followed by helium. Most of the rest of the elements in the universe were created in the stars, and are much less abundant. However, a few other nuclei besides hydrogen and ^4He emerged from the nucleosynthesis epoch, and these isotopes have important cosmological implications. To study these reactions in detail, cosmologists use computer models of nuclear reactions to predict the abundances of nuclei created immediately after the big bang. The constraints include such parameters as the density of nucleons and photons, the availability of neutrons, and the probability of occurrence of a given nuclear reaction. The probabilities of various reactions are known to very good accuracy, so the results of these models can be used with confidence to predict conditions during the nucleosynthesis epoch, provided that we can measure the abundances that actually exist. One important marker remaining from the epoch is the deuterium abundance. The precise abundance of primordial deuterium depends *very* sensitively upon the conditions in the universe during nucleosynthesis, especially the overall baryon density. The denser the universe, the less deuterium survives from the early nuclear reactions. If we can measure the amount of deuterium present today, accounting for destruction processes, we can use this measurement to derive the density of the universe. Like all cosmological observations, this is not an easy observation to make, but recent data have been sufficiently good to provide important constraints on the matter density in the early universe.

Most of the deuterium in the universe was created in the big bang, but it can be destroyed fairly easily within stars. Therefore, in order to measure the primordial abundance of deuterium, we must look for matter that has never passed through a star and is relatively uncontaminated by any subsequent nuclear activity. Measurements of deuterium abundances have been made in the atmosphere of Jupiter, in the local interstellar medium, and in the spectra of clouds of intergalactic gas that are illuminated by light from distant quasars.¹⁵ All of these determinations are quite consistent with one another, as astronomical data go, and give an abundance of $D/H \approx 1-4 \times 10^{-5}$, by mass, for the cosmological

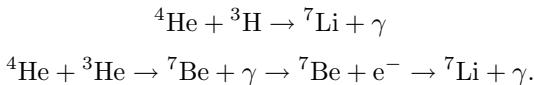
The deuterium abundance is a measure of cosmic density

¹⁵There is deuterium in the Earth's oceans, but deuterium is 10 times more abundant in the oceans than in the cosmos as a whole. Deuterium is favored over ordinary hydrogen to form water molecules. In the very early Earth, most of the deuterium became bound into water molecules and was thus unable to escape from the atmosphere, so its Earthly abundance cannot tell us anything about the early universe.

ratio of deuterium to hydrogen. The deuterium abundance places a limit on the density of baryons in the universe, since only baryons participate in the nuclear reactions that create it. However, deuterium alone sets only an upper limit to this quantity, because deuterium is so readily destroyed in stars.

Another rare species, ^3He , is a bit less sensitive to density, but its primordial abundance also drops off as the baryon density increases. This isotope of helium is fairly resistant to destruction outside of stellar interiors, and measurement of present-day ^3He abundances yield density estimates that are consistent with the results of the direct measurements of deuterium.

The isotope of lithium ^7Li was also produced in the big bang. Lithium in the early universe was produced in reactions such as



This isotope can also be created in some stellar events, as well as by cosmic rays. Moreover, ^7Li is easily destroyed at moderate temperatures, even those in the atmosphere of the Sun. It was never expected that observations of ^7Li would be able to tell us anything about the big bang, but from measurements first taken in the 1980s it became clear that careful observations could detect this isotope in the cool atmospheres of some very old stars. From such measurements, the primordial abundance of ^7Li could be inferred to be ${}^7\text{Li}/\text{H} \simeq 10^{-10}$. This is the abundance predicted for quite reasonable assumptions about the big bang. If the measurement truly indicates the primordial ^7Li , it represents a powerful vindication of the accuracy of the standard model of the early universe.

Taken together, the synthesis of the light elements in the early universe places a fairly stringent constraint upon the total density of baryons in the cosmos. Nucleosynthesis is almost entirely controlled by the temperature of the universe, and by the ratio of neutrons to protons. Model-dependent factors, such as the expansion rate and the geometry, influence nucleosynthesis only indirectly; such factors affect the cosmic time at which the universe reaches the appropriate temperatures, as well as control the density of nucleons and the neutron to proton ratio at the initiation of fusion. The major limiting factor to nucleosynthesis is the neutron, since all nuclei beyond hydrogen must contain at least one neutron. The more neutrons that decay before combining with protons, the smaller the abundance of heavier nuclei; this in turn depends upon factors such as the expansion rate. The earlier fusion begins, the more neutrons are available for the construction of heavier nuclei. Density, which is a function of the expansion rate, determines the reaction rates for both creation and destruction of nuclei. Deuterium is particularly sensitive to the density, since it is so easily destroyed if it interacts with other particles; thus the higher the density, the smaller the abundance of deuterium.

All the cosmological effects, as well as all the complications of nuclear and particle physics, must be taken into account if we wish to compute

Other light elements

Nucleosynthesis and baryon density

The baryon density of the universe

very precise values for the abundance of helium and the other heavier nuclei produced in the big bang. The result is a set of curves of the predicted abundance for each nucleus as a function of the density of the universe. If we superimpose the observational limits on a plot of these curves, we can determine whether there is overlap of the measurements of the primordial abundances of the different nuclei created in the big bang. The overlap of the measured abundances shows us the range in which the density of the universe is permitted to lie. If there were no overlap, then our model would be inconsistent in some way, or some of the data must contain errors that had not been taken into account. When we carry out this exercise, we find that the present baryonic density of the universe is very tightly constrained; it cannot be far from $\rho_b \approx 10^{-28} \text{ kg m}^{-3}$.

The raw density in baryons is not of much interest, however; we must compare the measured density of baryons to the critical density. That density can be written as a density parameter, that is, it can be expressed as a fraction of the critical density, $\Omega_b = \rho_b/\rho_c$. Of course, the definition of a density parameter includes the Hubble constant; when we compare the measured density to the critical density, then we are implicitly dividing by a factor of H^2 (*cf.* equation 11.14). Cosmologists prefer to state the density limits from nucleosynthesis in a form that is independent of H_0 . They define a new parameter h , where

$$h \equiv H_0/100, \quad (12.9)$$

and then write

$$\Omega_b \approx 0.019h^{-2} \quad (12.10)$$

as a mean estimate from the available isotope abundance data. If we use $h = 0.72$, corresponding to the current best estimate for H_0 , we get $\Omega_b \approx 0.04$, far below the critical value.

The density obtained from the light-element abundances would be equal to the *total* mass density of the present universe only if most or all of the matter is in the form of baryons. This appears not to be the case; it seems that most of the mass in the universe is due to some sort of exotic particle. The nucleosynthesis limit can tell us only what fraction of the total is in baryons; it places no bounds whatsoever upon the mass density due to nonbaryonic matter. Current observations show that the density of the universe is greater than baryons can provide, and we must conclude that some other form of matter, which does not participate in nucleosynthesis, is present.

Interesting information can also be obtained from measurements of the primordial abundance of ${}^4\text{He}$. Carrying out such a measurement calls for some care. Although most of the present ${}^4\text{He}$ in the universe was created in the big bang, it can also be produced in stellar nuclear reactions. Hence merely measuring its abundance in stars such as the Sun, which formed from the debris of stellar explosions, would not be adequate. Galaxies that have few metals (in astronomical usage, all elements heavier than helium), and thus have experienced relatively little

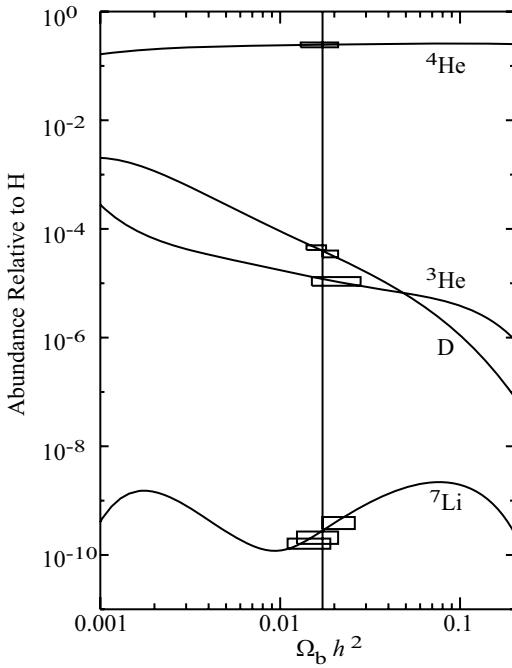


Fig. 12.8 Big bang nucleosynthesis abundances, specified as a fraction of hydrogen abundance, as a function of the baryon density parameter multiplied by the square of the Hubble parameter $h = H_0/100$. The solid curves are the predicted values, and the boxes indicate various measurements of the cosmic abundances of light-element isotopes. The vertical line is the best fit for this data set, a value of $\Omega_b h^2 = 0.019$. The rapid change in deuterium abundance as a function of density makes it a sensitive measure of the density parameter. (Adapted from Izotov et al., 1999.)

stellar nucleosynthesis, will not have created much additional ${}^4\text{He}$. Careful observation of such galaxies yields an abundance of between 0.22 and 0.26, by mass, for the fraction of primordial helium relative to hydrogen.

As it happens, the abundance of helium helps hardly at all to pin down the value of Ω_b . For helium, the higher density of a closed model, which would tend to increase the rate of fusion compared to the lower-density open model, is mostly counterbalanced by the slower expansion rate of a closed model, leading to later onset of nucleosynthesis and hence fewer available neutrons by the time the temperature has dropped to the point at which nuclei can survive. What is remarkable, however, is that the abundance of ${}^4\text{He}$ *can* be used to restrict the number of species of neutrinos. This is possible because the number of neutrino species affects both the expansion rate during nucleosynthesis, and also the temperature at which the ratio of neutrons to protons freezes. The more neutrino species, the faster the expansion and the earlier the freezing of the neutron-to-proton ratio, which results in more neutrons relative to protons. Both faster expansion and, especially, earlier freezeout tend to increase the production of ${}^4\text{He}$. Therefore, the more neutrino species, the greater the final abundance of ${}^4\text{He}$. The best estimate for the cosmic ratio of ${}^4\text{He}$ to hydrogen is close to 0.24 by mass. This, together with the observed abundances of deuterium and lithium, indicates that there are three species of neutrinos, which happens to be the number that has been detected. By the same token, the ${}^4\text{He}$ abundance must be at least near 0.23, since a lower value would require that only two species of neutrino exist, and we know of three. There are also theoretical reasons

Primordial helium abundances constrain the number of neutrino species

from particle physics to expect that exactly three species of neutrino should exist; thus the concordance of the helium data is comforting.

The observations of helium, deuterium, and lithium abundances show remarkable agreement with the predictions from the big bang model of cosmology, independently indicating very similar numbers for various parameters that enter into the model. This consistency cannot be taken lightly, and we must be careful before we start to tinker with such success. We conclude that the simplest model of the thermal history of the universe outlined here is quite successful at explaining the present abundances of light elements and their isotopes. On the other hand, the model is so constrained by the observations that the discovery of anything very much out of the ordinary, such as another species of neutrino, could force us to abandon the model. This is, of course, the hallmark of a scientific theory: it is falsifiable. There are still many open questions that address important issues in modern cosmology, some of which will be discussed in the following chapters. Nevertheless, the standard model of big bang cosmology must be regarded as a great achievement.

From light to dark: the end of the radiation era

At the end of the nucleosynthesis epoch, the universe contained photons, neutrinos, and ordinary matter in a state called a *plasma*, consisting of electrons, protons (hydrogen nuclei), ^4He nuclei (approximately 25% by mass), and traces of such light nuclei as deuterium, ^3He , and lithium. A plasma, in this context, is matter that is ionized, meaning that the negatively charged electrons are separated from the positively charged nuclei, and both the nuclei and the electrons act as independent particles. Essentially, the temperature is sufficiently high that the electrons have too much energy, and thus are moving too fast, to be captured by the electrostatic attraction of the nuclei. The cosmic plasma was still in thermal equilibrium with the photons at this point, meaning that the ordinary matter freely interacted with the photons and the same characteristic temperature applied to both. Most importantly, the free electrons scattered the photons randomly in all directions. With the matter density still quite high, a photon could not have traveled far before encountering a free electron and being diverted from its original path; light emitted from any point would quickly end up traveling in random directions. It would have been impossible to see very far through this dense plasma; during the radiation era the universe was opaque, like an incomprehensibly hot fog.

Radiation and matter densities were comparable during the epoch of equal densities

As the universe continued to expand, the temperature dropped in proportion to the scale factor. The mass density also fell, decreasing as the cube of the scale factor, while the radiation energy density continued to diminish as the fourth power of the scale factor. Although the radiation dominated the universe at the beginning, its energy density dwindled more quickly than did the density of matter, and hence eventu-

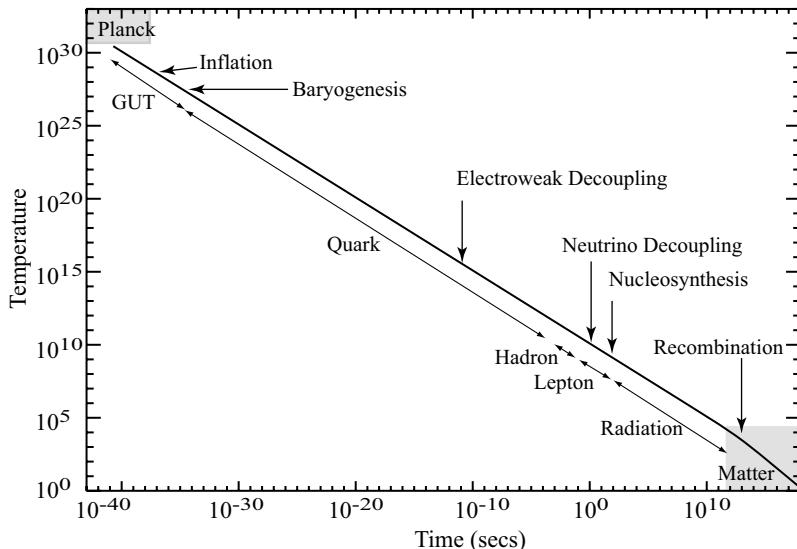


Fig. 12.9 Great moments in the history of the universe. Important events and epochs are shown along a line indicating radiation temperature versus time.

ally the matter became more important. The point of equality between the mass density and the radiation energy density occurred sometime around 10^{12} s. The time during which the matter and radiation densities were comparable can be called the **equal density epoch**. During this interval, it is a bit more complicated to derive the scale factor as a function of time in a closed form; we will not write down the formula, as it is not particularly illuminating on its own. It can be surmised that during this epoch the scale factor varied between $t^{1/2}$ and $t^{2/3}$, with the expansion occurring more like $t^{1/2}$ (characteristic of a radiation dominated universe) near the beginning of the epoch, and making a smooth transition to $t^{2/3}$ (characteristic of a matter dominated universe) by its end.

The epoch of equal density did not persist for long, because expansion was still quite rapid at this early stage. An important process began about this time, however. During the radiation era, the constant interplay between matter and radiation ensured that the plasma remained mostly smooth and homogeneous. But once the tight coupling was lost, it became possible for the matter to bunch together into clumps. These clumps are believed to be the seeds for the structures that formed later in the universe, namely the galaxies, clusters of galaxies, and any other great agglomerations that are present. Thus this was the time of **structure formation**, a phenomenon that will be discussed in more detail in Chapter 15.

After the epoch of equal density, the radiation energy density soon became negligible in determining the overall gravitation, and the universe entered the matter dominated era, which has lasted till the present day. The next great landmark was reached when the temperature cooled to approximately 3000 K, sometime around 10^{13} s. Below this temperature, electrons are no longer moving fast enough to escape from the

The universe becomes transparent

Great moments in history

Epoch	Time (s)	Major events
Planck	$0-10^{-43}$	All forces unified
GUT	$10^{-43}-10^{-35}$	Baryogenesis
Inflation	$\sim 10^{-37}$	Exponential increase in R
Quark	$10^{-35}-10^{-6}$	Universe of fundamental particles
Electroweak	$\sim 10^{-11}$	Weak and EM force decouple
Hadron	$10^{-6}-10^{-4}$	Matter excess frozen
Neutrino decoupling	~ 1	Background neutrinos go free
Lepton	$10^{-4}-10$	Proton/neutron ratio frozen
Nucleosynthesis	~ 100	Light atomic nuclei formed
Radiation	$10-10^{12}$	Scale factor goes as $t^{1/2}$
Recombination	$\sim 10^{13}$	Universe becomes transparent
Matter	10^{12} -now	Galaxies, stars, life

electric fields of the nuclei; the conditions then became suitable for most of the free electrons to be captured by the protons to form hydrogen atoms. This occurrence is known as **recombination**.¹⁶ After recombination, very few free electrons remained to scatter photons, so the photons streamed freely through the universe. This had two effects. First, radiation and matter ceased to be in thermal equilibrium. From this point onward, the radiation would simply cool with the expanding universe, with little regard for what the matter was doing. Second, the universe became transparent; it finally became possible to see for great distances. The edge of the fog, known as the *surface of last scattering*, is what we actually observe when we look back through space and time to the cosmic background radiation. This is why detailed observations of the CBR are so important; we are actually observing the state of the universe as it was only a short time after the big bang. The details of these observations will be discussed in Chapter 14.

The fate of a universe

How will the universe end?

We have examined the dramatic beginning of our universe; let us turn our attention to the question of its end. There are two general classes of possible ending: with a bang or a whimper. What would these two different fates be like? The spherical standard universe ends in a big crunch, the return to a very dense state such as was present at the beginning. The gravitational attraction of the matter in the universe is sufficiently great that it eventually halts, and then reverses, the expan-

¹⁶Since there were no hydrogen atoms previously, they must have simply combined for the first time; nevertheless, recombination is the standard terminology.

sion. If we lived in such a universe, the first sign of the contraction would be that relatively nearby galaxies would no longer show redshifts. Instead, we would observe only blueshifts in these galaxies as the collapse began. As galaxies fell together, the transition from redshift to blueshift would move to greater and greater distances; that is, to larger and larger lookback times. At first glance, it might seem that the evolution of the universe would run in reverse. However, conditions during the collapse will not be identical to those during the expansion because changes have occurred over the lifetime of such a universe. Inchoate matter at the beginning organized itself into stars, and the stars into galaxies. Plenty of hydrogen existed to make new stars; the galaxies were the lively places that we know today. But as time passes, the gas is used up; the stars cease to shine, with massive stars dying first in violent explosions while low-mass stars slowly burn out as cinder-like white dwarfs. At the end, the contents of the closed universe would be dominated by stellar ashes such as dead white dwarfs, neutron stars, and black holes. In fusing hydrogen to heavier elements, the stars converted mass into energy. In collapsing to black holes, the smooth, initially uniform, space-time continuum developed numerous singularities at the centers of these holes. These conditions, quite different from those prevailing during the expansion, are carried into the collapse and the new overall singularity.

The big crunch

The future history of the universe is quite different if it ends in a whimper, as is the case for flat, open, and accelerating universes. These universes exist forever; time begins but never ends. Yet as time increases into the infinite future, the average matter and energy densities drop toward zero, and all temperatures decrease toward zero. Stars burn out, white dwarfs cool as much as they can. Black holes, drawn together by gravitational radiation, merge with each other and with other, less exotic, remnants. But even the largest black holes, over a fantastically huge but not infinite time, decay via Hawking radiation into particles and photons. In some theories of particle physics the proton, the basic constituent of baryonic matter, turns out to be unstable, albeit with an enormous lifetime. If protons decay, all baryonic matter will ultimately break down into more fundamental forms. Eventually, it will no longer be possible to extract useful energy from anything; the universe will become ever colder and increasingly disorganized. These universes evolve toward a **heat death**, a state of minimum temperature and maximum disorder, or **entropy**.

A dark future

It is characteristic of the universe that its entropy increases as it evolves. What, then, is entropy? There are several ways in which to think about it, but fundamentally, it is a quantitative measure of the disorder of a system. For example, suppose that all the air in a room were confined into a small volume by some contrivance. That state would be reasonably well ordered, as gases go. Now suppose the air is released; it fills up the room and the molecules mix more or less randomly. As it expands, the initially compressed air could do some work, such as turning a turbine. But once the air is thoroughly mixed, it can no longer do any work. The motions of the molecules are random, not systematic, and

Entropy defined

cannot become sufficiently organized to turn a turbine. We say that this new state, with the air spread throughout the room, has more disorder, that is, higher entropy, than the old. As another example, the gasoline that fuels a car is in a relatively ordered state, containing substantial available chemical energy. When the engine burns the gasoline, the vapor is converted into various gases, some of which push the pistons that ultimately cause the wheels to turn; after each stroke, these combustion products are vented through the exhaust system. The waste gases are in a much more disordered state than was the original gasoline, so the final entropy is much higher than the initial. This process cannot be run in reverse; the exhaust cannot be collected and pushed backwards through the engine to create gasoline. Energy has been released at the expense of creating quite a lot of entropy.

In a closed system entropy cannot decrease

Thermodynamics is the science of energy and entropy, and the **second law of thermodynamics** states that the total entropy at best remains the same, and usually increases, in any process. It is possible to reduce entropy and create order locally by means of the expenditure of energy; but such a process always results in an overall increase in the entropy of the universe. The human body is an ordered system, but to exist it must utilize considerable energy from food. The energy is available in low-entropy chemical form (for example, as sugar molecules). Most of that energy turns into higher-entropy random heat and is radiated from the body, unavailable for further use. The body maintains its relatively ordered state at the expense of producing greater disorder in its environment.

Nearly all the energy on the Earth is, or was, ultimately provided by the Sun. Energy from the Sun drives the atmospheric and oceanic motions that control the Earth's climate. Photosynthetic plants and bacteria capture energy from the Sun's light; some of this energy they use to manufacture organic compounds for their own use, while the rest is stored in chemical form. When other organisms consume plants, the stored energy becomes available to them. Fossil fuels such as petroleum and coal are the remains of ancient plants; these fuels are, in a sense, frozen sunshine. On average, the Earth neither gains nor loses energy over periods of roughly a year or longer. Energy not used to maintain the *status quo* is reradiated in the form of heat and returned to space. A device that uses a temperature (and entropy) differential to perform work is a heat engine. It takes energy from some high-temperature source and exhausts higher-entropy heat at a lower temperature. Thus the Earth system, including the biosphere, is a huge heat engine. The Earth receives usable (lower entropy) energy from the Sun; its systems (atmosphere, ocean, biosphere, etc.) then extract work, which goes into purposes that include sustaining low-entropy entities such as living creatures; the unused energy is then re-emitted in the form of high-entropy, lower-temperature heat. The high temperature of the radiation beaming from the Sun onto the lower-temperature Earth is what makes life possible.

The Sun is a ball of gravitationally compressed gas that contracted from a diffuse, cold cloud of interstellar gas and dust. How can we reconcile the presence of stars, which form spontaneously, with the second law of thermodynamics? After all, stars might seem to be more ordered than a swirling gas cloud. However, gravity complicates the picture. Stars do not consume energy as they form. Whether they ever reach a state of nuclear fusion or not, proto-stars release energy as they contract; if they do ignite nuclear reactions, considerably more energy is produced. Our previous examples have suggested that release of energy tends to be accompanied by an *increase* in entropy. In a gravitating system, the higher-entropy states are those that are contracted. The more clumped the matter, the higher the entropy. The ultimate is the black hole, which is maximally contracted and is in a state of very high entropy. This may seem counterintuitive, because it is directly opposite to the case of a gas in which gravity is negligible; in such a gas, the more diffuse states have higher entropy. But when gravity is present, the diffuse gas has the potential by contracting to release its gravitational energy, and possibly to perform work, a capacity that is reduced the more the gas clumps. Hence a star actually has higher entropy than the gas from which it formed, and we can easily account for the spontaneous creation of stars. The clumping of gas into stars, of stars into galaxies, and of galaxies into clusters obeys the second law: entropy increases.

Entropy increases in gravitational contraction

What is the entropy of the entire universe? One component of the total entropy is measured by counting photons. At the beginning of the universe, both matter and antimatter existed in near-equal amounts. Most of the photons in the universe resulted from the mutual annihilation of this matter and antimatter. The photons are in thermal equilibrium; that is, they represent a blackbody. Since they are all in equilibrium, there are no variations in the temperature of the photons that could be used to perform work. If it had been possible to separate the matter and antimatter into different boxes in the very early universe, the matter–antimatter reaction could be used now for some application such as powering spaceships, and there would be fewer photons in the universe. The entropy of the universe would then also be lower. Another component of the universe’s entropy is in the mass of the black holes. The formation of black holes over the lifetime of the universe is another manifestation of the inexorable increase in entropy.

Entropy and the second law of thermodynamics play a key role in defining the direction of the **arrow of time**. The laws of mechanics, and of special and general relativity, and even of quantum mechanics, show no obvious asymmetry in time; they do not disclose why time has a preferred direction. Yet a broken bottle never spontaneously reassembles itself. The air in a room does not abruptly coalesce into one corner, leaving a vacuum in the rest of the room. Left to their own devices, things evolve from order into disorder. So it is for the universe. Why would a contracting universe not be simply an expanding universe running backwards in time? The direction of time appears to be defined by the direction of increasing entropy, and this is independent of the

The arrow of time

Cyclic models

behavior of the scale factor; the universe always moves from an ordered state to a state of greater disorder. The universe winds down.

Speculating on the ultimate fate of the universe is a particularly stimulating pastime. In the case of eternal universes, entropy increases and the opportunities for further extraction of useful energy diminish. The universe becomes dominated by high-entropy photons, all at some incredibly low temperature, and there is not enough energy available to be converted into appreciable amounts of work; these universes simply fade away. Such a heat death does not lie in the future for the closed universe, but there too entropy may have important implications. It is a popular notion that if the universe returned to a big crunch, it might bounce to a new big bang and start again. This would be a cyclic universe, a sort of modified steady state model. The universe continues into the indefinite past and the unending future, but it has individual manifestations, each separated by the infinite crunch of a singularity. If it turned out that the universe could somehow pass through the singularity at the big crunch and rise from its own ashes, it seems probable that, unlike the phoenix, it would not return as simply a younger version of itself. The big crunch would be in a state of higher entropy than was the big bang. What if the universe *tunneled* through the singularity, reemerging in another big bang? If the new universe remembered its entropy, then the next universe would have more photons, fewer particles, and higher entropy than the old, and, with less matter density, would expand to a larger size than in its previous incarnation. After some number of cycles, the matter density contained in the new universe would be too low for the formation of galaxies and stars. There is a greater difficulty inherent in the cyclic model, however. The big crunch would consist of many merging black holes, separate singularities coming together into one final universal singularity. A singularity created from the mergers of many black holes is quite different from the initial singularity out of which our universe emerged. If the universe were to reemerge in a new manifestation, could it find its way back to the smooth state such as apparently existed during our big bang? This is a question for which we have no answer, although it seems unlikely. In the absence of a quantum theory of gravity, we cannot know much of how singularities behave. Even so, there is really nothing to suggest that a big crunch could produce another big bang.

So long as there is no definitive observation that determines the fate of the universe, we are free to state an emotional preference for its ultimate destiny. Would it be better to end in fire, or in ice? Some people, including some cosmologists, opted for a steady state model rather than face the implications of an end to the universe as we know it. Some seek a compromise in a closed universe that goes through endless cycles. But, others counter, is eternal repetition really preferable? Nothing ever really changes in the steady state model; there is simply an endless repetition of events. In an infinite steady state universe, presumably there is an infinite number of Earths, with all possible histories, but still

very much the same. In the standard models, there is an end. If we pass this way but once, we must make the most of it.

Chapter Summary

The history of the universe can be divided into many epochs, depending on which constituent was most important. Today the matter density of the universe dominates its gravity completely; thus we say that we live in the matter era. However, conditions were not always as they are today. As we look backward in time toward $t = 0$, the universe becomes increasingly hot. Density also rises ever higher as we look toward earlier times, but the variation with scale factor differs for matter density and for radiation energy density, with matter varying as the inverse cube of the scale factor, while radiation energy density varies as the inverse fourth power of the scale factor. Hence there was a time in the past when radiation was more important than ordinary matter in determining the evolution of the universe. The interval of domination by radiation is called the radiation era.

Temperature is a measure of energy, and Einstein's equation $E = mc^2$ tells us that energy and matter are equivalent. At sufficiently high temperatures, particles with large mass can be created, along with their antiparticles, from pure energy. The temperature also influenced the behavior of the fundamental forces of nature during the earliest intervals of the universe's history. As we approach $t = 0$, we encounter increasingly unfamiliar epochs, dominated by different physics and different particles. The earliest was the Planck epoch, during which all four fundamental forces were unified and particles as we know them could not have existed. Next followed the unified epoch, when gravity had decoupled but the other three forces remained unified. During this epoch, the scale factor may have undergone a tremendous exponential expansion. The small excess of matter that makes up the universe today must have been created during the unified epoch, by a process still not completely understood. As the temperature continued to fall, the universe traversed the quark epoch, the hadron epoch, the lepton epoch, and the epoch of nucleosynthesis.

The first atomic nuclei formed during the nucleosynthesis epoch. Most of the helium in the universe was created from the primordial neutrons and protons by the time the nucleosynthesis epoch ended, scarcely three minutes after the big bang. A few other trace isotopes, specifically deuterium (heavy hydrogen) and ^7Li , were also cre-

ated; their abundances depend sensitively upon the density of the universe during this time. If the universe were too dense, then most of the deuterium would have fused into helium. Only in a low-density universe can the deuterium survive. The major factor controlling the ultimate densities of helium and deuterium is the abundance of neutrons. The more neutrons that decay before combining with protons, the smaller the abundances of elements heavier than hydrogen. The availability of neutrons in turn depends on the expansion rate as well as on the cosmic matter density. Comparing the observed densities of the primordial isotopes to those computed from models and translating the results into the Ω density parameter, gives $\Omega_b = 0.019h^2$, where h represents the Hubble constant divided by $100 \text{ km s}^{-1} \text{ Mpc}^{-1}$. Nucleosynthesis limits indicate only the density of baryons, because only baryons participate in nuclear reactions. Hence we must conclude that the universe contains less than the critical density of baryons.

After nucleosynthesis, the universe continued to cool uneventfully for roughly a million years. During this time the ordinary matter consisted of a hot plasma of nuclei and electrons. The free electrons made the plasma opaque; a photon of radiation could not have traveled far before being scattered. However, once the universe cooled to approximately 3000 K, the electrons no longer moved fast enough to escape the attraction of the nuclei and atoms formed, an event known as recombination. At this point the universe became transparent to radiation. The last moment at which the universe was opaque forms the surface of last scattering; it represents the effective edge of the universe since we cannot see into the dense plasma that existed prior to recombination. Once the radiation was able to stream freely through the universe, matter and radiation lost the tight coupling that had bound them since the beginning. Henceforth matter and radiation evolved almost entirely independently. The photons that filled the universe at the surface of last scattering make up the CBR today, but now their energy is mostly in the microwave band.

At a time of $\sim 10^{12}$ s the matter density and the energy density were equally important. This marked the onset of the epoch of structure formation. The seeds of structure

formation may have been planted much earlier, during the GUT epoch, but the tight coupling between radiation and matter prevented the density perturbations from evolving. As matter and radiation began to evolve separately, density perturbations could begin to grow. This process continued after recombination as the most over-dense regions collapsed gravitationally, forming galaxies and clusters of galaxies.

The evolution of the universe is marked by an increase in entropy. Entropy is a quantitative measure of disorder.

The second law of thermodynamics states that in any physical process, the overall entropy at best remains the same and usually increases. The arrow of time seems to point in the direction of increasing entropy, beginning from the low-entropy singularity of the big bang. The future of the cosmos is not yet certain. While it could be closed and end in a big crunch, the most likely fate is that it will expand forever. In such a universe, temperatures cool toward absolute zero and entropy increases to a state called heat death.

Key Term Definitions

nucleosynthesis The process by which nuclear reactions produce the various elements of the Periodic Table.

thermal equilibrium A state in which energy is equally distributed among all particles, and all the statistical properties of the particles can be described by a single parameter, the temperature.

matter era The epoch of the universe, lasting from approximately the time of recombination until the present, during which the energy density of radiation is negligible in determining the overall gravitational field of the universe, and the mass-energy density of matter is dominant.

radiation era The epoch in the history of the universe, lasting from the big bang until approximately the time of recombination, during which the energy density of radiation controlled the gravity of the cosmos.

pair production The creation of a particle and its antiparticle from some form of energy, such as photons.

quantum mechanics The theory that describes the behavior of the very small, such as molecules, atoms, and subatomic particles. It is spectacularly successful at explaining experimental data, but gravity cannot yet be made to fit within the theory.

Compton wavelength The quantum wavelength of a particle with a highly relativistic velocity.

graviton A hypothetical massless boson that is the carrier of the gravitational force.

symmetry The property under which some quantity does not change when certain attributes, such as spatial location, time, rotation, and so forth, vary.

spontaneous symmetry breaking The loss of symmetry that causes fundamental forces to become distinguishable. In most theories, this occurs in the early universe when the temperature becomes low enough that the different energy scales of the various forces become important.

Planck epoch The epoch from the beginning of the universe until the Planck time. Very little is known about this interval, although probably all four fundamental forces were united.

unified epoch That interval in the early history of the universe when three of the four fundamental forces, the strong and weak interactions and the electromagnetic force, were unified.

grand unified theory A member of a class of theories that seek to explain the unification of the strong, weak, and electromagnetic forces.

inflation A period of exponential increase in the scale factor, due to a nonzero vacuum energy density, that occurs early in the history of the universe in certain cosmological models.

hadron A class of particles which participate in the strong interaction. Hadrons consist of those particles (baryons, mesons) which are composed of quarks.

lepton A member of a class of fermionic particles that do not participate in the strong interaction. The best-known lepton is the electron.

- quark** One of the six fundamental particles that make up hadrons.
- baryon** A fermionic particle consisting of three quarks. The most important baryons are the proton and the neutron.
- baryogenesis** The creation of matter in excess of antimatter in the early universe. Only the relatively few unmatched matter particles survived to make up all subsequent structures.
- quark epoch** The interval in the early universe during which quarks were unconfined in hadrons and were dominant.
- gluon** A hypothetical particle that binds quarks together into hadrons.
- electroweak interaction** The unified electromagnetic and weak forces. Also called the electroweak force.
- hadron epoch** That interval in the early history of the universe after the quarks had condensed into hadrons, and before the temperature dropped below the threshold temperature for protons.
- lepton epoch** The interval in the early history of the universe during which leptons dominated.
- nucleosynthesis epoch** The interval in the early history of the universe when helium was created, along with traces of a few other light element isotopes.
- deuterium** An isotope of hydrogen whose nucleus contains one proton and one neutron.
- equal density epoch** That interval in the early history of the universe when the gravitational contributions of matter and radiation were approximately equal.
- structure formation** The process by which the large-scale structure in the universe, namely the galaxies, galaxy clusters, and superclusters, developed from small density perturbations in the early universe.
- recombination** The moment in the early universe when the temperature became sufficiently low that free electrons could no longer overcome the electrostatic attraction of the hydrogen nuclei and were captured to form atomic hydrogen. When this occurred the universe became transparent.
- heat death** The fate of the open universe models in which the temperature drops toward zero, stars die out, black holes evaporate from Hawking radiation, entropy increases, and no further energy is available for any physical processes.
- entropy** A quantitative measure of the disorder of a system. The greater the disorder, the higher the entropy.
- second law of thermodynamics** The law that states that the entropy of a closed system always increases or at best remains the same in any process.
- arrow of time** The direction, apparently inviolable, of the “flow” of time that distinguishes the past from the future.

Review Questions

- (12.1) What might have been some reasons that the predictions of Gamow and his collaborators were ignored for so many years?
- (12.2) Supporters of the steady state model argued that the gradual creation of matter out of nothing in empty space is philosophically preferable to the big bang, which features creation of everything out of nothing at $t = 0$. Do you agree? How are these two types of creation physically and logically distinct?
- (12.3) Consider the universe at a redshift of $z = 2$. What is the average matter density in the universe then compared to now? What is the average radiation energy density then compared to now? What is the temperature of the background radiation then? Suppose the energy density in matter, ρc^2 , is now 2000 times greater than the energy density in radiation. At what redshift did matter and radiation have equal energy densities?
- (12.4) Describe what is meant by thermal equilibrium. What simplifications does this state of equilibrium allow us to make in our description of the early universe? What does equilibrium imply about the rates of reactions, as well as the numbers of photons and massive particles?
- (12.5) The electron has a mass of 9.11×10^{-31} kg. What is the threshold temperature in kelvin for an electron

to be created from gamma rays, assuming thermal equilibrium? What is the Compton wavelength of the electron?

- (12.6) (More challenging.) At the Planck time, $t = 10^{-43}$ s, the temperature is 10^{32} K. The formula for the scale factor during the radiation dominated era can be written

$$R(t) = R_P(t/t_P)^{1/2},$$

where R_P is the Planck length and t_P is the Planck time. Assuming that this formula holds for the duration of the radiation era, (a) compute the temperature at cosmic time $t = 10^{-33}$ s (the approximate time of the grand unified theory symmetry breaking). (b) Compute the redshift from now back to this GUT time. (c) Compute the Hubble length at this GUT time ($c = 3 \times 10^8$ m s $^{-1}$). (12.11)

- (12.7) What would happen if there were significant accumulations of antimatter anywhere in the universe? (12.13) Why do such antimatter accumulations seem unlikely to exist? What significance does this conclusion hold for particle physics in the unified epoch? (12.14)
- (12.8) Why did the lepton era end with more protons than neutrons in the universe?
- (12.9) What can observations of the cosmic deuterium abundance tell us about the big bang? Why is deuterium a particularly good nuclide to use for

this purpose? List at least two complications that make the measurement of deuterium an uncertain proposition.

Characterize the following eras in the history of the universe: matter era, radiation era, nucleosynthesis epoch, lepton epoch, hadron epoch. For each of these eras, what were the main components of the universe, and what were the most important phenomena that occurred? What events ended each era?

If a fourth species of neutrino were discovered, what might it mean for the standard hot big bang model?

What was the era of recombination? What happened then, and why is it significant? What is meant by the *surface of last scattering*? How does this surface limit the information we can obtain about the universe?

What is the connection between the early universe and the cosmic background radiation?

What is the ultimate fate of the closed universe? Of the open? Of the flat? Explain what is meant by the *heat death* of the universe. What fundamental law of physics predicts such an ultimate state? Why is a heat death an inevitable consequence of this law for certain universes? Why does this same law appear to make a cyclic universe unlikely?

Testing the Models

13

Key Terms:

Observatory, *n.* A place where astronomers conjecture away the guesses of their predecessors.

Ambrose Bierce, *The Devil's Dictionary*

The parameters of the universe

Current observations show that the universe is consistent with the cosmological principle; on the largest scales it is truly homogeneous and isotropic. We know that the universe is expanding; its scale is increasing with time. Because the universe is isotropic and homogeneous, we can characterize that increase in scale by a scale factor $R(t)$ that is a function only of time. A model of the universe is a description of this function. The models are derived from the most basic set of assumptions: general relativity describes the behavior of gravity on the large scale, and the universe conforms to the cosmological principle. Add to this the observational fact that the universe is expanding, and we conclude that the universe began a finite amount of time in the past, and was once much more compact, dense, and hot than it is now. This hot, dense, initial phase of the universe is known as the big bang.

In the standard models, the universe contains matter, and the gravitational force exerted by that matter causes the expansion rate to slow over time. The matter content of the universe also determines its overall geometry. A low-density universe is hyperbolic and expands forever. A high-density universe is spherical and will eventually recollapse into a big crunch. Between these two possibilities is the flat universe, a dividing line marked by a critical density value. The flat universe has a Euclidean geometry and expands forever, although at a rate that approaches zero as time increases. These standard models are derived from the simplest possible set of assumptions for a nonempty, homogeneous, and isotropic universe. If we go beyond the standard models by adding a cosmological constant, many more possible evolutionary paths become possible for the universe. When Λ is present, the universe might not have begun with a bang. Even stranger histories are possible. With the appropriate positive value of Λ , we can create a model that is infinitely old and spent the first half of its infinite existence in a state of collapse. In such a model, the Λ -force causes a turnaround at a minimum size,

Models of the universe

- parameters of the universe
- dark matter
- emission distance
- reception distance
- redshift-distance relation
- standard candle
- angular size
- kinematical method
- mass-to-light ratio
- dynamical method
- dark halo
- virial theorem
- cold dark matter
- hot dark matter
- vacuum energy
- quintessence
- dark energy
- concordance model

after which the universe begins expanding again, heading back out to infinity. This is an inverted closed universe, in a sense. However, most of these models are without any compelling basis in observation. The two guiding observational facts are the cosmological redshifts indicating that the universe is expanding, and the substantial evidence that there was a big bang. A big bang is possible even with a cosmological constant, as long as matter dominates in the very early universe; thus the possible presence of a cosmological constant does not inherently contradict the physical evidence of the initial hot, dense state of the universe.

Even with the restriction to an isotropic, homogeneous universe demanded by the cosmological principle, and expansion and big bang initial conditions required by observations, we still find a plethora of models from which to choose. The task, then, is to determine which model best agrees with the actual universe. What determines how $R(t)$ changes with time? Cosmologists work with a convenient set of variables that characterize the evolution of the universe in terms of potentially measurable quantities. These are the **parameters of the universe**. The total number of parameters that cosmologists can consider is fairly extensive, but the most essential of these can be quickly summarized.

The first of these is the *Hubble constant*, H_0 ; the subscript indicates the value in the universe today. The Hubble constant describes the rate of expansion of the universe; it is similar to a velocity, measuring how rapidly the scale factor R changes with time. The Hubble constant also sets the basic time- and length-scales for the universe. Its inverse $1/H_0$ is the Hubble time, an estimate of the age of the universe, while c/H_0 is the Hubble length, a measure of the size of the observable universe. Mathematically, the Hubble constant is the rate of change of R with respect to time, divided by R ; that is, it is specified by \dot{R}/R .

The *curvature* of the universe, k , describes the geometry of space. Specifically, if the universe is hyperbolic, $k < 1$; if spherical, $k > 1$; and if flat, $k = 0$. In the standard models the overall geometry of the universe is determined by the matter-energy content of the universe, but with the addition of Λ there is greater freedom.

Because the evolution of the universe is determined by its contents, it is important to measure the overall matter-energy density and the value of Λ . It is convenient to measure these values using dimensionless parameters defined in terms of the critical density of the universe. The overall matter density of the universe is expressed in terms of the *mass density parameter*,

$$\Omega_M = 8\pi G \rho_0 / 3H_0^2. \quad (13.1)$$

Similarly, a Λ parameter can be defined as

$$\Omega_\Lambda = \Lambda c^2 / 3H_0^2. \quad (13.2)$$

The utility of these parameters can be seen by relating them to the geometry of the universe. Specifically, $\Omega_M + \Omega_\Lambda - 1 = kc^2/R_0^2H_0^2$. We can further define a new space curvature parameter,

$$\Omega_k = -kc^2/R_0^2H_0^2, \quad (13.3)$$

Measuring the parameters of the universe

The density parameters defined

to produce the particularly simple relationship

$$\Omega_M + \Omega_\Lambda + \Omega_k = 1. \quad (13.4)$$

Note that these parameters are defined in terms of their values at the current time, t_0 . Knowledge of any two of these parameters determines the third. For example, if $\Omega_\Lambda = 0$ and $\Omega_M < 1$, the matter density is low and the geometry is hyperbolic, but if $\Omega_M > 1$, density is high and the geometry is spherical. If the geometry is flat, then $\Omega_k = 0$ and the sum $\Omega_M + \Omega_\Lambda$ must be equal to unity. In its most general form, we can use the Ω parameters to account for everything that can contribute to the gravitational forces in the universe. For example, if radiation made a significant contribution to the universe today, we would include an Ω_r term. Furthermore, Ω_M can be broken into subcomponents. One component, defined as Ω_b , consists of ordinary matter, that is, matter composed of *baryons*; another component is composed of mass in the form of nonbaryonic **dark matter**. For the moment we shall refer to this parameter as Ω_{DM} .

The *deceleration parameter*, q_0 , tells us whether the rate of expansion is decelerating, with $q_0 > 0$, or is accelerating, with $q_0 < 0$. Mathematically, q_0 measures a second derivative of R with respect to time, that is, the rate of change of the rate of change of R ; it is defined as $q = -\ddot{R}/RH^2$. The value of q_0 is determined by the forces acting on the contents of the universe. The most important datum is whether the universe is decelerating or accelerating, that is, whether q_0 is positive or negative. Attractive gravitational forces from ordinary matter and energy produce decelerations and, left to themselves, result in $q_0 > 0$. However, a positive value of Einstein's cosmological constant Λ can produce an overall repulsive force. With a sufficiently large Λ term, it is possible that $q_0 < 0$. We can write q_0 in terms of the Ω parameters as

$$q_0 = \frac{1}{2}\Omega_M - \Omega_\Lambda. \quad (13.5)$$

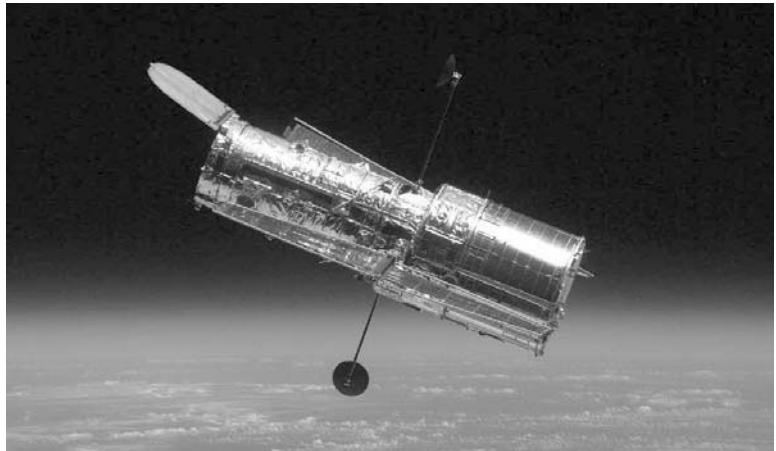
Therefore the universe is accelerating if $2\Omega_\Lambda > \Omega_M$.

A specific model predicts precise relationships among the various parameters, for example, H_0 , Ω_M (including both baryonic and dark matter), and Ω_Λ (from a cosmological constant). The determination of these parameters in turn predicts values for other quantities such as the geometry, the age of the universe t_0 , and q_0 . Knowledge of the exact values of these parameters would answer many questions about the history of the cosmos. In principle, these cosmological parameters are observable, and our program thus seems simple: measure as many parameters as possible, in as many ways as possible, and then compare the data to the predictions of various candidate models in order to deduce which model best describes our universe. Upon completion of this effort we would know whether the universe is open, flat, or closed, we would know its age to perhaps a few million years, we would know a great deal about its matter and energy content, and we would know much about its ultimate fate.

The deceleration parameter

An observational program for cosmology

Fig. 13.1 The *Hubble Space Telescope*. The telescope is designed to carry out observations in the visible and ultraviolet regions of the spectrum while in near-Earth orbit. The unprecedented resolution of the telescope is possible because it is above the Earth's atmosphere, which blurs and distorts ground-based observations. The primary mirror of the *HST* is 2.4 meters in diameter. (STScI/NASA.)



Fortunately, the immediate prospects for the success of such a program are quite good. Observations have recently improved dramatically, thanks to space missions such as the *Hubble Space Telescope (HST)*, the *Cosmic Background Explorer (COBE)*, and the *Wilkinson Microwave Anisotropy Probe (WMAP)*. The new generation of giant ground-based telescopes, such as the 10-meter *Keck* telescope in Hawaii, are also providing valuable new data. From at least the time of Galileo, history has shown that the introduction of new observational capabilities inevitably leads to significant progress, as new portions of the universe are opened to scrutiny for the first time. Although we still have far to go before any final statements can be made about the values of the important cosmological parameters, cosmologists now feel that we have entered a time in which precision measurements are feasible.

The Hubble constant and redshift–distance relations

The Hubble constant is a key component in determining the fate and structure of the universe. The value of the Hubble constant enters into the definitions of the Ω parameters and determines the Hubble time, $1/H_0$, which sets the scale for the age of the universe. Therefore, a good place to begin our observations is with a continuation of Hubble's original program of measuring redshifts and distances for a large ensemble of galaxies. The Hubble constant is difficult to determine because of the requirement of obtaining accurate distances. The systematic uncertainties inherent in various rungs of the distance ladder have led to a significant disparity in measured values of the Hubble constant, with values ranging historically from around 50 to 100 $\text{km s}^{-1} \text{ Mpc}^{-1}$. However, dramatic recent results have begun to narrow down the range of possible values. One of the most important developments was the completion of the *HST* key project to observe Cepheid variables in the Virgo Cluster of galax-

ies, the nearest large galaxy cluster. The research team, led by Wendy Freedman, was able to determine the distance to the Virgo Cluster with unprecedented accuracy. Once this benchmark was established, it was possible to obtain a better calibration of other standard distance indicators that can be used on ever more distant galaxies.¹ Combining all the latest results yields a current best value of $H_0 = 72 \pm 7 \text{ km s}^{-1} \text{ Mpc}^{-1}$. While the stated error in this measurement does not, of course, take into account *unknown* sources of error, astronomers have attempted to allow for systematic errors in their error estimates. In any case, this latest value is consistent with other measurements made since the 1990s, many of which have bounced around in the range of approximately 65 to 80 $\text{km s}^{-1} \text{ Mpc}^{-1}$. Perhaps it is not overly optimistic to suggest that we have finally closed in on the elusive value of H_0 .

The value of H_0 alone is not enough to specify the type of universe. Nevertheless, measurement of the change in H over time provides a way to begin to discriminate among models. The time dependence of H over longer intervals is observable because we are looking back in time as we look out into space. For example, if we determine the redshift z to a distant galaxy, we are measuring the value of $(1 + z) = R_0/R_{\text{then}}$ at a lookback time of $t_0 - t_{\text{then}}$, when the light from that galaxy first started on its way to us. Next we must measure the distance to the galaxy, but the expansion of the universe complicates somewhat the idea of a “distance to a galaxy.” At the time t_{then} when the galaxy emitted the light, it was at a certain distance, the **emission distance**, from us. When we receive that light today, the distance to that galaxy has increased by a factor of $(1 + z)$ owing to the Hubble expansion. This distance is called the **reception distance**. The reception distance provides a measure of distance at a specific instance of cosmic time, t_0 . These distances scale according to the rule for distances in an expanding universe, as given by the Robertson–Walker metric (10.6). It must be emphasized that *both* the emission and the reception distances depend individually upon the cosmological model, but their ratio is a simple function of the redshift:

$$\frac{D_{\text{now}}}{D_{\text{then}}} = \frac{R_0}{R_{\text{then}}} = 1 + z. \quad (13.6)$$

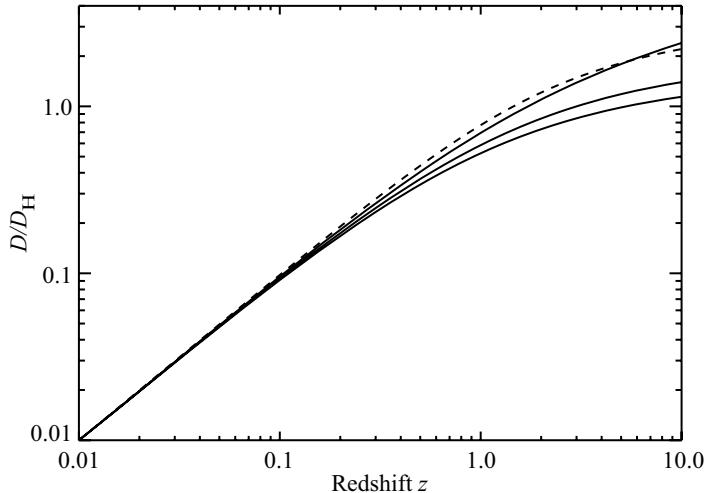
The emission distance is smaller than the reception distance by precisely the ratio of the scale factor then to the scale factor now. A galaxy at a redshift of $z = 2$ was one-third as far as its current distance from us when the light we receive today was emitted.

A collection of redshifts and distances provides a history of the scale factor $R(t)$. For example, a high-density universe decelerates more rapidly than does a low-density universe. Hence in a high-density universe, the scale factor was smaller more recently than in a low-density cosmos. Therefore, a given redshift corresponds to a smaller distance in a high-density universe in comparison to one of low density. The theoretical curve of $R(t)$ allows us to construct for each type of model a plot

Different measures of distance in an expanding universe

¹Some of these methods are described in Chapter 10.

Fig. 13.2 Reception distance, as a fraction of the Hubble radius, versus redshift for various cosmological models. From bottom to top the solid lines correspond to a closed standard model with $\Omega_M = 2$, a flat $\Omega_M = 1$ model, and a hyperbolic model with $\Omega_M = 0$. The dashed curve corresponds to a flat model with $\Omega_M = 0.3$ and $\Omega_\Lambda = 0.7$, currently regarded as giving the best agreement with observations. The models are indistinguishable until beyond a redshift of nearly 0.1. At higher redshifts, where measurements could most easily discriminate among the models, accurate data become increasingly difficult to obtain.



of reception distances as a function of redshift, the so-called **redshift-distance relation**. We can then assemble a set of curves of predicted redshift versus distance for each model universe. One such set of curves using reception distances is illustrated in Figure 13.2.

Obtaining distances from observations is somewhat tricky. Assuming we know a good **standard candle**, we can measure the apparent brightness as a function of redshift for a large sample of these standard objects. The apparent brightness of a standard candle is a proxy for distance, since in flat spatial geometry the brightness drops off as the distance squared, $b \sim L/D^2$. This computation is complicated somewhat if the universe has spherical or hyperbolic spatial geometry. If space is not flat, we require a geometry correction to relate the luminosity distance to the emission distance. The surface area of a sphere is larger or smaller than $4\pi r^2$ if the geometry of space is hyperbolic or spherical, respectively. Suffice it to say that for all of our candidate models we can derive for each redshift a luminosity distance that can be compared with observations.

Unfortunately, it is very difficult to distinguish among different models with observations. Although different models produce different curves on a theoretical plot of redshift versus the brightness of standard candles, the deviation of the curves from one another becomes significant only for large redshifts and correspondingly large distances. But high-redshift data are difficult to obtain, and systematic errors abound. The absorption of light by matter in intergalactic space and within our own Milky Way is one source of uncertainty. There are other, more subtle, complications as well. One of the most important potential sources of error is the possible evolution of the source with time. As we look further and further into space, galaxies and other potential standard candles are younger and younger. Their brightness may well change with time, rendering them quite unsuitable as a standard candle. The redshift itself may also change a galaxy's appearance. Astronomers can observe in only

Luminosity distance is affected by expansion and spatial geometry

Differences among models become significant only at large redshift

one band (visible, radio, etc.) at a time, but observations in different regions of the electromagnetic spectrum emphasize different attributes. For example, at higher and higher values of z , detectors sensitive to visible wavelengths will actually be observing light that was emitted, in the galaxy's own frame, from increasingly bluer portions of that galaxy's spectrum; the image will not show the normal *optical* appearance of a galaxy, but will be skewed toward objects bright in blue or ultraviolet light. Although this effect can be accommodated in observations, the various sources of uncertainty still tend to be larger than the spread produced by variations among models. Just where precision measurements become particularly important to discriminating between models, our ability to make such measurements becomes very poor.

One of the best of the standard candles, Cepheid variables, cannot be observed at high redshift. They are simply not bright enough; telescopes lack the necessary resolution to distinguish individual Cepheids at great distances. But Cepheids are not the only possible standard candle. The most useful standard candle at high redshifts is a particular type of supernova, the Type Ia, which results from the explosion of a white dwarf star. The first step in establishing these supernovae as standard candles was to calibrate their brightness by determining the luminosity through independent distance measurements. This is an area in which the *HST* results proved particularly valuable. The *HST* was able to observe Cepheids in relatively nearby galaxies in which a supernova had also once been seen. From such data, astronomers have determined that Type Ia supernovae have nearly the same maximum brightness and, more precisely, their peak brightness is directly related to the rate of decline of the light after the peak.

The Type Ia supernova is a cosmological distance indicator

The major advantage of Type Ia supernovae is that they are so bright that they can be seen at very great distances. Their disadvantages are that they are rare, they are short-lived, and they are seemingly random events. This creates some significant challenges. First, a supernova must be detected. Then astronomers must follow up quickly with observations on a large telescope while the fading light is still bright enough for detailed study. But time on major telescopes like the *Keck* telescope or the *HST* is in great demand, is fully allocated, and must be scheduled well in advance. How can observations be scheduled for supernovae of unknown location or time of explosion? The answer is that the *rate* is sufficiently well known that astronomers can compute how many galaxies must be observed in order to see a few supernovae each month. It is then possible to schedule follow-up observations for the supernovae that are statistically expected to be seen. It is rather like planning a press conference for the winner of an upcoming lottery. No one knows who will win, but it is very likely someone will. Such a project requires a systematic, organized search, and in the 1990s two teams of astronomers set out to do just that. The basic approach is to make long-exposure images of regions of the sky at intervals separated by approximately a month. The two images are compared to determine whether any changes occurred during that time. When a supernova is found, detailed follow-up ob-

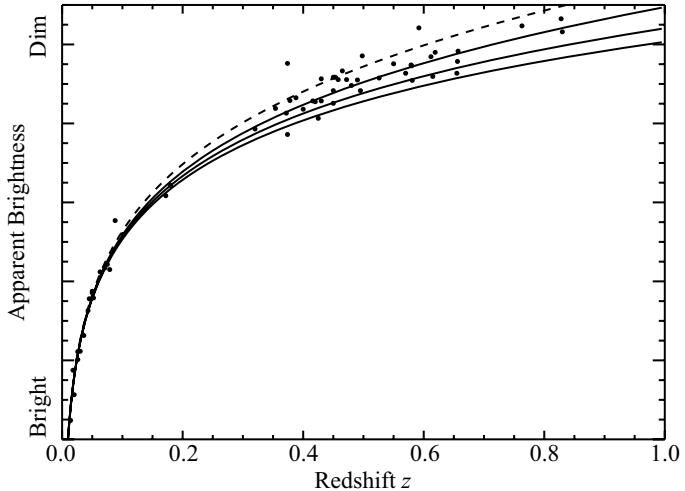


Fig. 13.3 The observed brightness (apparent magnitude) of Type Ia supernovae versus redshift, compared with different models of the universe. The three solid lines, from bottom to top, correspond to a closed ($\Omega_M = 2$), flat ($\Omega_M = 1$), and open, empty ($\Omega_M = 0$) models. The dashed line is an accelerating model with $\Omega_\Lambda = 1$. The basic result is that the distant supernovae are dimmer, that is, they are further away than would be expected for a standard model with $\Lambda = 0$. (Adapted from Perlmutter et al., 1999.)

servations are performed using the already scheduled telescope time, in order to observe the rate at which the supernova's luminosity declines. This determines the brightness of the supernova, which in turn gives the distance. In this way, data for supernovae at increasingly high redshift values have been gradually accumulated. By the time a few dozen supernovae had been seen, an amazing trend had already emerged: the data favor an *accelerating* universe. In a standard model, the deceleration parameter is given by $q_0 = \Omega_M/2$, and q_0 must be positive. More generally, $q_0 = \Omega_M/2 - \Omega_\Lambda$. In other words, a negative value for q_0 requires that the universe is accelerating and must contain a nonzero cosmological constant.

Figure 13.3 shows the apparent brightness of the observed supernovae as a function of redshift. The solid lines plot the expected curves for some standard models, while the dashed line corresponds to an empty, flat, de Sitter type universe ($\Omega_\Lambda = 1$). All models converge at low redshift; for small redshifts the observed supernovae provide a calibration of the Type Ia supernova's intrinsic luminosity. The observations have found that distant supernovae are dimmer on average than would be expected in any of the standard models. The best fit to the data yields a model with $\Omega_M \approx 0.3$, $\Omega_\Lambda \approx 0.7$, and a universe that is flat.

The data contain considerable scatter, and there are many sources of systematic error that could influence the results. For example, a supernova's apparent brightness could be decreased by a greater amount than that due only to distance if a significant quantity of dust lies along the line of sight. Moreover, there may be more intrinsic variation in peak supernova brightness than we currently appreciate, particularly for high-redshift supernovae that occurred billions of years in the past. Some of these concerns were alleviated by the discovery of a particularly distant supernova by the *HST*. Astronomers found the supernova in a long exposure of a portion of the sky known as the Hubble Deep Field. Comparing data taken in 1997 with data from 1995, they found

Distant supernovae appear too dim for a decelerating universe

a change in brightness in a galaxy at redshift $z = 1.7$, due to the most distant supernova seen to date. This new supernova lies precisely on the redshift–magnitude curve predicted by the other supernova data. If intervening dust were causing the supernovae to appear dim, thereby making it seem that the universe was accelerating, then this new supernova would look even dimmer according to a simple, known relationship. However, at a redshift of 1.7 the universe has not begun to accelerate yet, so the supernova actually looked a bit brighter than it would have if the effect were due to dust. *HST* has since found additional high-redshift supernovae, and the data are consistent with a universe that contains a significant cosmological constant Λ and began accelerating around $z = 0.5$. Systematic errors due to dust or intrinsic changes in supernova brightness now seem to be ruled out.

The prospects for an even greater understanding of the universe from supernova data are very good in the near future. All that is needed is to observe an ever larger number of supernovae. Astronomers have proposed a new space mission for this purpose, the *SuperNova Acceleration Probe* or *SNAP*. This experiment would consist of an orbiting telescope dedicated to locating and observing supernovae at redshifts between $z = 0.1$ and $z = 1.7$. This experiment could observe thousands of supernovae, compared to the few dozen that have been found so far. Such a huge sample could determine the acceleration of the universe with impressive precision.

The age of the universe

As surprising as the supernova results were, astronomers had begun to suspect that Λ is not zero for other reasons. As measurements of the Hubble constant began to converge toward the relatively high value of $72 \text{ km s}^{-1} \text{ Mpc}^{-1}$, cosmologists faced a knotty theoretical problem in the form of the age of the universe. The Hubble constant implies a Hubble time and, for any decelerating universe, the Hubble time is the upper limit to the actual age of the universe. For $H_0 = 72 \text{ km s}^{-1} \text{ Mpc}^{-1}$ the Hubble time is 13.9 billion years. Each specific standard model predicts an age for the universe that is some fraction of the Hubble time. The age of the flat standard universe is two thirds the Hubble time,

$$t_{(\text{Einstein--de Sitter})} = \frac{2}{3H_0} = \frac{2}{3}t_H, \quad (13.7)$$

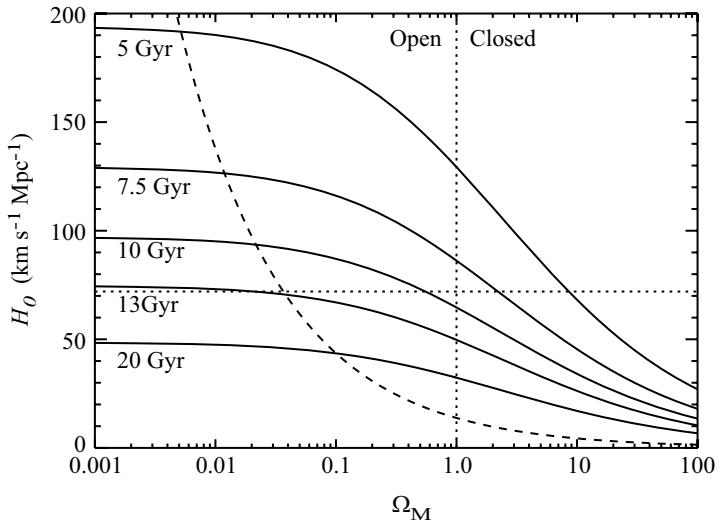
which for the above Hubble time is merely 9.3 billion years. For the open model, the age lies between $\frac{2}{3}t_H$ and t_H , depending upon the density; the lower the density, the closer the actual age is to the Hubble time. In the spherical model, the age is *less than* $2/3 t_H$.

It is difficult, if not impossible, to find the exact age of the universe independent of a model. The best we can do is to obtain the ages of various constituents of the universe. Obviously, the universe must be at least as old as its oldest components. Well-established radioactive dating techniques indicate that the solar system is about 4.5 billion years

A supernova at $z = 1.7$ was a fortunate discovery by HST

Age and the Hubble time

Fig. 13.4 The age of the universe for the standard models. The plot shows the relationship between the Hubble parameter, the age of the universe, and the density parameter Ω_M for models with $\Lambda = 0$. The solid lines trace the value of H_0 and Ω_M for universes with the indicated ages. The dashed line corresponds to $\Omega_b h^2 = 0.019$, the value obtained from studies of big bang nucleosynthesis. The dotted vertical line is the critical density $\Omega_M = 1$, and the dotted horizontal line is $H_0 = 72 \text{ km s}^{-1} \text{ Mpc}^{-1}$, the best current value for the Hubble constant. With this value of H_0 only an open standard model is old enough to accommodate globular cluster ages, and even then only if the oldest stars are less than 13 billion years in age.



Globular cluster stars provide a lower limit for the age of the universe

old; this sets an absolute lower bound on the age of the universe. Models of the Sun are consistent with this, indicating that it is approximately 4.6 billion years old. Of all astronomical objects, stars are probably the best understood; their ages can be estimated with good confidence from stellar models.² We can seek a lower bound for the age of the universe by determining the age of its most ancient stars. The oldest stars that we know are located in globular clusters; their inferred maximum ages range from perhaps 10 to around 18 billion years, with the stellar modelers favoring something toward 12–14 billion years. If that is so, then stellar ages are a fairly severe constraint on cosmological models. This is illustrated in Figure 13.4, which shows the relationship between cosmological age, H_0 , and the density parameter Ω for the standard ($\Lambda = 0$) models. If the Hubble constant is greater than $75 \text{ km s}^{-1} \text{ Mpc}^{-1}$, the flat and spherical universe models are simply too young to account for such objects. Even the empty, open model has little time to spare, and it cannot accommodate the upper range of stellar ages. In fact, the cosmological imperative has been a reason why some stellar modelers actively sought mechanisms to reduce the derived ages of the globular clusters. Otherwise, there would probably be near agreement that the globular clusters are 12–14 billion years old. However, if the universe is *accelerating* due to a cosmological constant, then the age of the universe can be *greater* than the Hubble time, and all geometries are acceptable. Turning the argument around, if we determine that the universe is flat, then the age of the universe requires that Λ be present.

Historically, the age crisis is something that has come and gone. When Hubble first determined a value for the parameter now named for him, the number he reported corresponded to a universe with a Hubble time of only 2 billion years, a value much less than the known age of the Earth!

The age crisis

²Chapter 5 discusses some of the ways in which stellar ages can be determined.

This is one of the reasons that the Lemaître model experienced a period of interest, and why the steady state cosmology enjoyed popularity for as long as it did. The systematic errors that caused this overestimate of H_0 were not corrected until the 1950s. It is possible that some unknown systematic error may yet be distorting our modern results. However, many recent measurements, using different approaches, have independently obtained $H_0 \sim 72 \text{ km s}^{-1} \text{ Mpc}^{-1}$, and studies of the cosmic background radiation³ are fully consistent with the implications of the Type Ia supernova results. Combining all the various results statistically, the best age of the universe is found to be approximately 13.7 billion years. Cosmologists' confidence is growing that the true age of the universe is finally becoming established.

The geometry of space

In the standard models the density parameter completely determines the geometry of space, Ω_k . However, if Λ is not zero, the models have additional freedom, and the independent measurement of cosmological parameters becomes crucial. In the standard models Ω_Λ is zero; measurement of Ω_M , the critical density in ordinary matter, determines whether space is spherical, hyperbolic, or flat. It is exceedingly difficult to determine Ω_Λ through direct measurement; its value, however, could be obtained indirectly if we could measure the matter density of the universe and the geometry of space independently. Determination of these quantities by similar methods would not do, since the errors would then tend to go in the same direction, distorting the result of the subtraction. With the possibility of a cosmological constant raised by the stellar ages and the supernova data, it becomes especially important to determine the geometry of space from independent measurements. How might this be done?

An interesting method of measuring the geometry of the universe exploits the dependence of apparent size upon geometry and distance. In ordinary Euclidean geometry the **angular size**, that is, the angle occupied upon the sky, of a *known* length is a direct measure of distance, since the angular size θ , along with the distance D to the object, gives the actual size; that is, the angular size is simply inversely proportional to the distance. If we symbolize the proper length of the object by ℓ , then the Euclidean formula is just $\ell = \theta D$. But expansion and non-Euclidean geometry introduce new effects. When we speak of distances in an expanding universe, we must make a distinction among different ways of describing distance. This complication arises from the fact that looking out in space means looking back in time. The quasar light that we see today was emitted eons ago, when the quasar was much closer to our galaxy. At first, as we look to greater and greater distances, the galaxies' angular sizes become smaller and smaller as usual; they are

Measuring triangles in space

³These studies will be discussed in more detail in Chapter 14.

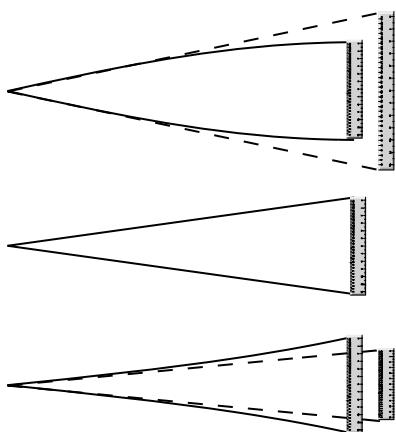


Fig. 13.5 The effect of spatial geometry on angular size. A standard length appears larger in spherical geometry (top) and smaller in hyperbolic geometry (bottom), compared to flat geometry (middle).

farther and farther away, after all. But eventually a point is reached at which the lookback time becomes important. At significant lookback times, the universe was appreciably smaller. Very distant galaxies were much closer to us at the time of their light's emission; in fact, for very large redshifts, they would have practically loomed over us. Since we see such a galaxy as it was then, it appears larger on the sky than it would had it been located at that distance all along. Therefore, beyond some turnaround point at which emission distances become small, the apparent size of a galaxy actually *increases* as a function of redshift. The amplitude of this effect depends on the scale factor $R(t)$.

The angular size-distance relationship also depends upon the geometry directly. The apparent angular size of some standard length ℓ would differ for the three types of geometry even if space were not expanding. The different geometries predict different relationships between geometrical quantities such as the radius and circumference of a circle, or the radius and volume of a sphere. In particular, the sum of the interior angles of a triangle depends upon geometry. Consider a triangle with two sides of equal length D and a third side of length ℓ , where D is much greater than ℓ . In flat geometry the angle between the equal sides is $\theta = \ell/D$. In hyperbolic geometry the angle would be smaller than this, whereas in spherical geometry that angle would be larger; hence in hyperbolic space an object would seem further away, while in spherical space it would appear closer. Thus the relationship between angular size and distance provides a means to determine geometry directly from observations, through its dependence upon $R(t)$ and upon Ω_k . We can plot theoretical curves of θ as a function of redshift for the different models (Figure 13.6). If a standard length existed, we could observe it and overlay the data on such a plot to determine the geometry. Several tests have been carried out, using various standard lengths such as the size of the largest spirals in a galaxy cluster, or the lengths of radio jets emerging from remote radio galaxies. One of the difficulties with the test of angular size is that galaxies and radio jets are not really very good standard lengths. Also, as can be seen from Figure 13.6, data must be collected at relatively high redshifts before significant differences in the standard models become apparent.

Recent experiments using the angular size-distance test have proved more successful. Rather than using galaxies, radio jets, or other intrinsically variable objects as standard lengths, these new observations use something quite different: the angular size of the small temperature fluctuations in the cosmic background radiation.⁴ Theory predicts the physical extent of these regions of very slightly hotter or cooler gas in the early universe. It is then possible to predict how large these regions will appear on the sky for the different geometries. The results obtained from the *WMAP* spacecraft are consistent with a flat universe, $\Omega_k = 0$. This outcome may not be very exotic, but at least it means that the

Fluctuations in the CBR provide a standard length at a redshift $z = 1000$

⁴See Chapter 14.

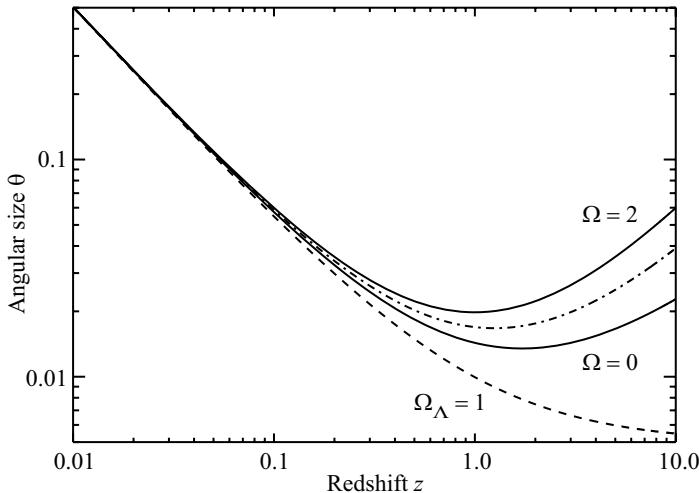


Fig. 13.6 Angular size θ as a function of redshift z for an object of a given proper length. The lines are theoretical curves for different cosmological models. From top downward the curves correspond to a closed $\Omega_M = 2$ universe (solid line), flat $\Omega_M = 1$ (dot-dashed line), open empty $\Omega = 0$ (solid line), and de Sitter $\Omega_\Lambda = 1$ (dashed line) models. Angular size is a simple decreasing function of redshift at low z , but starts to increase with distance for some models beyond a redshift of around $z = 1$.

Euclidean geometry most of us learn in school applies to the universe as a whole.

Before leaving this topic we will describe another method of measuring geometry that does not require a specific standard length. If galaxies are more or less equally spaced throughout the universe, they could be used to measure the size of circles, or the volume of spheres, with a given radius. In this way it would be possible to determine the geometry of the universe without a standard length, and without measuring redshifts at all, simply by counting the number of sources at a given optical or radio apparent brightness for some area of the sky. If galaxies were identical, then the apparent brightness of any galaxy's image would be immediately related to its distance. A count of the number of galaxies at each brightness (distance) would thus yield an estimate of the volume of space at that distance. The different geometries predict different number counts, with the hyperbolic geometry giving more sources while the closed geometry gives fewer, as illustrated in Figure 13.7. In principle, this method could determine the curvature without the need for accurate distance measurements. In practice, this approach is rather complicated.

Unfortunately for astronomers, galaxies are not all the same, but vary in their intrinsic brightness; they may also change in brightness over their lifespans. In order to compute the theoretical curves required for comparison of the data with the model, we must make assumptions about the nature of the sources we are studying; specifically, we must either assume that the galaxies do not change over the large lookback times involved, or we must develop a detailed model to describe the evolution of these sources. There is little doubt that the first approach rests on a very dubious assumption. Looking back in time, we can see changes in the appearance and apparent properties of galaxies and quasars. For example, there are no, or at best very few, nearby quasars, but many at high redshifts. Normal galaxies also change with time, growing brighter

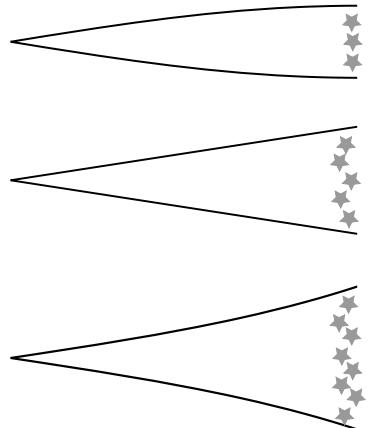


Fig. 13.7 Number counts of galaxies can distinguish among the different geometries. If galaxies were distributed equally throughout space, different geometries would show different numbers of galaxies as a function of distance. Spherical geometry has the least volume, hyperbolic the most.

Number counts of galaxies could disclose the geometry of the universe

as stars form, or as they collide and merge, and becoming dim as their stars age and burn out. How can we develop a model for this complicated evolution? At the present, we cannot account for all observed phenomena, but must build our best model based upon what we know. If we can develop a good model for the distribution of the inherent brightness of a large population of galaxies, and if we can somehow account for possible changes in their brightness over time, then we can apply corrections for these effects to our counts of number at a given brightness. The brightness of any individual galaxy may not be a function of its distance alone, but with the corrections for galactic variability, it is possible to arrive at an overall figure that stands as a proxy for the distance *distribution* of the galaxies under study.

Another weakness of this approach is that sources, such as galaxies or quasars, are not equally spaced throughout the universe, but cluster. Further corrections must be applied in order to separate the effects of clustering from the effects of geometry. Yet another confounding effect is the possibility that the spatial distribution of galaxies, not just their separation, might change over the history of the universe due to factors unrelated to the geometry of space. For example, the number of galaxies per unit volume might increase as galaxies are formed, or their numbers might decrease if they merge. All of these complications have an effect upon the number count that is far greater than that due to the geometry of space itself. Despite this, the initial attempts at number count observations made an important contribution to the history of cosmology. Even the mostly inconclusive early results were inconsistent with the perfect cosmological principle and the steady state model; they showed that the universe and its contents *are* evolving, although they could not pin down a geometry. In fact, although number counts and surveys of galaxies started off as a way to gauge the geometry of the universe, astronomers now use them to gain valuable information about the evolution and clustering statistics of galaxies.⁵

The mass of the universe

One of the most important descriptors of the universe is the matter density parameter, Ω_M . The density of the universe is connected to other major cosmological parameters, such as the Hubble constant, the deceleration parameter, and the geometry. The crucial question has always been whether Ω_M is greater than, less than, or equal to the critical value of unity. If, as observations from the *WMAP* satellite indicate, the geometry of the universe is flat and Ω_k is zero, either $\Omega_M = 1$ and $\Lambda = 0$, or the sum $\Omega_M + \Omega_\Lambda = 1$. How much gravitating mass does the universe contain, and how much Λ ? How might we weigh the universe to determine these quantities?

⁵This is discussed in greater detail in Chapter 15.

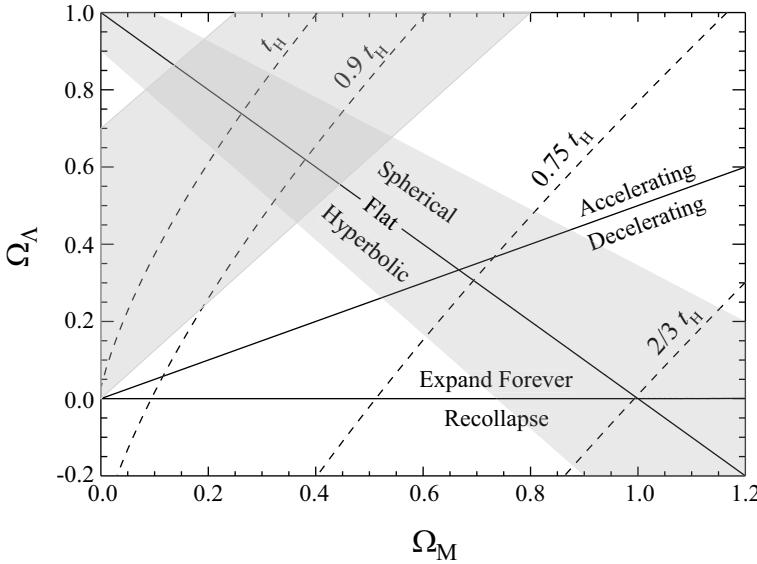


Fig. 13.8 The cosmological parameter space of Ω_Λ versus Ω_M . The solid lines indicate dividing points between geometry, models that accelerate or decelerate, and models that expand forever or experience a big crunch. The dashed lines indicate the age of the model in terms of the Hubble time. The shaded region rising from left to right shows the results and the range of uncertainty from the supernova surveys, while the shaded region rising from right to left comes from the CBR measurements indicating a flat universe.

The methods we have so far discussed measure the density parameter Ω_M indirectly, by determining the overall expansion rate and structure of the universe and matching the observations to a model. These techniques are called **kinematical methods**. Their major drawback is that they require precise determinations of distance and motion over a wide range of redshift, and such data are very difficult to obtain. For example, the Type Ia supernova measurements can be used to seek the best-fitting curve to the ensemble of data in terms of independent Ω_M and Ω_Λ parameters. The supernova measurements support the existence of a cosmological constant, but there are a number of combinations of parameters and models that could accommodate the results. An illustration of this is given by Figure 13.8, which plots Ω_Λ against Ω_M . Solid lines on the plot divide up the parameter space into models that accelerate from those that decelerate, those with hyperbolic geometry and those with spherical. Also overlaid on the plot are dashed lines indicating the age of the universe in terms of the Hubble time.

Although the results from the supernova studies and the observations of the CBR seem to indicate an accelerating flat universe, for such a significant result it is imperative to confirm this conclusion independently. Measuring the ordinary matter density parameter Ω_M directly is one way to accomplish this. The value of Ω_M can, in principle, be determined by summing up the total mass contained within some representative volume of the universe. The procedure seems straightforward: select a volume of space, count up the galaxies within it, and multiply by the mass per galaxy. A galaxy represents a significant, localized density enhancement, perhaps by a factor of 10^5 relative to the overall cosmic density, so there may be some question as to whether the galaxies are representative of the overall matter distribution in the cosmos. However, it seems to be

Plot of Ω_Λ versus Ω_M

The mass-to-light ratio in galaxies

the case that galaxies contain most of the *visible* matter in the universe. If the visible matter tracks the total matter, we can weigh the galaxies and then average their densities over all space, in order to obtain Ω_M .

How can a galaxy or a cluster of galaxies be massed? The most easily observed feature of any galaxy is its light. For nearly all galaxies, most of the light comes from stars, with some contribution from glowing regions of hot gas. Thus the amount and spectrum of the light from a galaxy gives us an estimate of the number and type of stars. A knowledge of the distribution of stars enables us to estimate the mass of the luminous matter, since we know quite accurately the mass of a star with a given temperature and luminosity. What we must determine is something called the **mass-to-light ratio** of the galaxy, M/L . For example, if a galaxy consisted of nothing but a collection of stars identical to our Sun, then in units of solar mass and solar luminosity, M/L would equal unity. We would then need only measure the total luminosity of this sun-galaxy to obtain its total mass. Of course, in a real galaxy there are stars that are more massive than the Sun, and many more times as luminous, but these stars are relatively rare. There are also stars less massive and substantially less luminous than the Sun, but they are much more abundant. Taking all these factors into consideration, a representative distribution of stars in a typical spiral galaxy has a mass-to-light ratio of a little more than one. The Sun, it seems, represents a reasonably good average.

But stars are not the only constituent of galaxies; obtaining M/L is not as simple as summing the masses of all the stars. Although only luminous matter is directly visible, all the mass in a galaxy, whether in stars, gas, or some mysterious exotic particle, will make itself known through its gravitational influence. We can use Newton's laws and the observed gravitational interactions of galaxies, gas, and stars to infer the mass density required to produce such motions. Methods of determining Ω_M that depend upon observing the dynamic interactions of stars and galaxies are known as **dynamical methods**. For example, if the galaxy is rotating, then we can exploit Kepler's laws to measure the total mass in the galaxy. Combining the rule for centrifugal force with Newton's law of universal gravitation shows that a body in a Keplerian orbit at radius r from the center of the galaxy obeys

$$GM(r) = v^2 r, \quad (13.8)$$

Obtaining mass through orbital velocities

where $M(r)$ is the total mass within radius r . Spiral galaxies rotate, making this technique mainly suited to them. A plot of the rotational velocity versus radius is the *rotation curve*; Kepler's law thus enables us to use the rotation curve to compute the mass at any radius. In particular, one of the characteristics of a spiral galaxy is a fairly sharp dropoff in the light distribution at some distance from its center. If we measure the rotational velocity at this radius, we can determine the total mass within the luminous portion of the galaxy.

If the luminous matter accounted for all or most of the gravity, then beyond the outermost circle of stars $M(r)$ would become constant, and

orbital velocities v would then decrease with increasing r according to equation (13.8). Furthermore, if essentially all the mass of a galaxy were contained within its luminous regions, then we should see a clear trend of decreasing velocities as we approach the edge, even before reaching the outlying stars. Although it becomes more difficult to track rotation curves where we cannot see any stars, there are ways in which orbital velocities beyond the edge of the main disk of a galaxy can be studied. In most cases, it is possible to find stray stars and globular clusters outside the luminous disk of very nearby spirals. Another valuable technique, still applicable only to spirals, is to observe the radio emissions of neutral hydrogen. This emission can be seen well beyond the visible edge of the disk, which immediately indicates that at least some gas surrounds typical spirals beyond the boundaries of their luminous matter. The radio output of these gas envelopes is relatively weak, however, suggesting that the gas is rather tenuous, and thus perhaps not itself a very significant mass contributor.

The interesting conclusion from observations to date is that for most spiral galaxies, there is no evidence of *any* decrease of orbital velocity with radius. The velocity $v(r)$ is roughly constant, or even increases, as r increases (Figure 13.9). By equation (13.8) it follows that $M(r)$ must also continue to increase with radius, despite the dearth of visible stars. Thus we cannot say with any certainty at all what is the total mass of a typical spiral galaxy; the galaxies must be surrounded by **dark halos** of unseen matter. Some evidence from rotation curves indicates that the halos of spiral galaxies may be spherical, and may extend to a considerably larger radius than that of the visible galaxy. It is from dynamical methods that we learn that the ratio of mass to luminosity in spiral galaxies is approximately

$$(M/L)_{\text{spiral}} \sim 10 \text{ to } 30 M_{\odot}/L_{\odot}, \quad (13.9)$$

where M_{\odot} is the mass of the Sun and L_{\odot} is the solar luminosity. This ratio provides the most compelling evidence that spiral galaxies contain substantial dark matter.

So far we have discussed only spiral galaxies, but elliptical galaxies are also a major constituent of the universe. It is more difficult to obtain estimates of the total mass of elliptical galaxies than it is for spirals since, by and large, ellipticals do not show any systematic rotation. The stellar motions within the ellipticals are still governed by gravity, however. One approach is to measure the *dispersion* of the velocities of stars, that is, the extent of the range of velocities about the mean velocity. We can then make use of a simple formula called the **virial theorem**. This is a statistical result for gravitating systems that relates a measurement of the velocity dispersion to the mean gravitational field. The virial theorem is somewhat similar in concept to hydrostatic equilibrium in a ball of gas, such as a star. In a star, the atoms must move fast enough (that is, have an adequately high temperature) to generate sufficient pressure to resist gravitational collapse. In a galaxy the stars do not collide, so there is no gas pressure, but they still must keep moving with

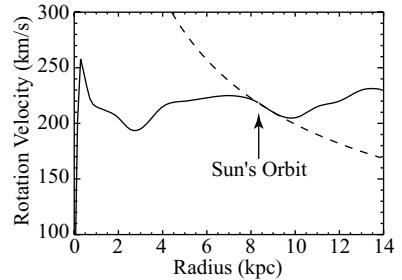


Fig. 13.9 Rotation curve for the Milky Way. The dashed line corresponds to the Keplerian velocity curve that would be observed if most of the mass of the Galaxy interior to the Sun's orbit were concentrated at the center. The fact that the rotation velocity does not drop off means that the total mass continues to increase with radius.

The mass-to-light ratio for spiral galaxies

The mass-to-light ratio for elliptical galaxies

an average velocity high enough to avoid gravitational collapse. This statistical rule provides an estimate of the collective mass of the system. Such measurements indicate that the M/L for most elliptical galaxies is comparable to that for spirals, $M/L \sim 20M_{\odot}/L_{\odot}$.

Such a large mass-to-light ratio implies that most of the matter of the galaxy is much less luminous than is the Sun. It is clear that galaxies contain considerably more mass than the luminous matter can explain; this is sometimes known as the *missing mass* problem. However, the existence of dark matter need not be startling. We are, after all, aware of *some* such nonluminous matter, specifically, the Earth and its fellow planets. Of course, the dark members of our solar system contribute negligible mass compared to the Sun. Generic small-massed planets are often called “jupiters,” after the most massive planet in our solar system. Like the planet Jupiter, they have some substantial mass but do not shine. We have scant knowledge of the distribution of jupiters throughout the Galaxy, but such objects most likely contribute very little to the total mass, particularly if they occur only in conjunction with a normal star. Even the archetypal Jupiter has only 0.001 of the mass of the Sun.

If planets cannot account for the missing mass, might it be found in previously unsuspected multitudes of faint stars? The luminosity of a star on the main sequence is approximately proportional to the third power of its mass.⁶ The rapid increase of luminosity with mass means that bright, massive stars are responsible for most of the light output of a galaxy; but stars of greater than about two solar masses are extremely rare. Because of this, lower-mass stars, the *red dwarfs*, contribute disproportionately to increasing the mass-to-light ratio. Such stars are, in fact, quite abundant; the Sun is actually near the upper end of the luminosity range of the relatively common stars. How can we determine whether the mass-to-light ratio for galaxies can be explained by these objects? One approach would be to study all the stars near the Sun, where even low-luminosity stars such as red dwarfs should be detectable. Such observations have shown that the median mass for nearby stars is approximately one third the solar mass, and the overall mass-to-light ratio in the solar neighborhood is roughly

$$(M/L)_{\text{nearby}} \sim 3M_{\odot}/L_{\odot}. \quad (13.10)$$

If the stars around the Sun are representative, as we believe them to be, it does not appear that dim, normal stars alone can explain the observed M/L for galaxies as a whole.

Brown dwarf stars

Faint stars and planets are not the only way to increase M/L ; there exists a population of small substellar objects with masses that lie between those of stars and planets. These objects are called *brown dwarfs*, and they have perhaps a few percent of the mass of the Sun, but very little luminosity. Could there be a huge population of these small, substellar

⁶See Chapter 5 for a discussion of the mass–luminosity relationship for main sequence stars.

objects? From the general observation that smaller objects are more common, it might be naively expected that many of these brown dwarfs should exist. Such stars are subluminous even for their tiny mass; they could, in sufficient numbers, make a very significant contribution to the large M/L . They are very difficult to observe, precisely because they are so dim, but determining their density is quite important to cosmology. Their existence has been confirmed; the first example of a brown dwarf, the faint companion of a star called Gliese 229, was detected in 1995 and imaged by the *HST*.⁷ Although many more examples of brown dwarfs have been discovered since then, there is little evidence for an extraordinary abundance of such objects. Star surveys have shown an unexpected cutoff to the mass distribution of self-luminous stars. Stellar theory had predicted that the minimum mass of an object that could ignite thermonuclear fusion at its core, and thus could be defined to be a star, should be about $0.08M_{\odot}$. Recent surveys instituted to look for brown dwarfs, as well as for very-low-mass main sequence stars, have turned up far fewer than had been predicted. It now seems that the total number of these small-massed stars and brown dwarfs may be comparable to the number of more ordinary stars in the Galaxy but, because their masses are so small, they make a minor contribution to the total galactic mass.

Brown dwarfs are stillborn stars; they never quite became hot enough in their cores to ignite nuclear fusion. But what about defunct stars, those that have used up their fuel? Some of the unseen matter could take the form of dim, compact stellar cinders such as white dwarfs; their luminosity, at a given mass, is quite low. Neutron stars would be even better, for they emit almost no light and, unless they beam as a pulsar directly toward our line of sight, they are nearly invisible. Best of all would be massive black holes, which could contribute a fairly large amount to the total mass without increasing the luminosity at all. The mass-to-light ratio of such dark objects would be essentially infinite! There is evidence that a realistic distribution of these stellar remnants, along with the main sequence stars, can explain the mass-to-light ratio of $M/L \sim 10M_{\odot}/L_{\odot}$ that is observed in the disks of spiral galaxies. However, the *overall* M/L for galaxies is still several times larger. White dwarfs seem to be too scarce to explain these larger ratios of mass to light; neutron stars and black-hole candidates are thought to be even rarer.

What about matter that has not assembled itself into either stars or planets? We know that spiral galaxies contain a great deal of interstellar gas and dust, which contributes to the $M/L \sim 10$ observed in the disk of the Milky Way. Unfortunately, measurements of the total mass of such matter fall far short of what would be required to make much of a contribution toward a ratio of $M/L \sim 30$. In our Galaxy it seems that the M/L of stars, gas, dust, and dwarfs can account for most of the mass-to-light ratio within the narrow galactic disk itself, but overall we are still missing quite a bit of mass.

Stellar remnants

⁷This object is shown in Figure 5.2.

Very little mass in the universe is luminous

For cosmological purposes we must convert the observed mass-to-light measurements into an Ω_M value. We begin with an estimate of the portion of the mass density of the universe contributed by visible matter. Overall, estimates of luminous mass from light output and distribution indicate that

$$\Omega_{LM} \simeq 0.005 \text{ to } 0.01. \quad (13.11)$$

This result shows clearly that luminous matter is a long way from closing the universe, and this conclusion is essentially independent of the value of the Hubble constant.

But spirals have massive dark halos; how much do they contribute to the total density of the universe? If, as it appears, a galaxy's rotation velocity becomes constant at large radius, then from equation (13.8), $M(r) \propto r$. Hence the total density is proportional to r^{-2} . Since the radii estimated for distant galaxies are dependent on the value of the Hubble constant, the halo mass density scales in the same way as the critical density. The resulting dynamical estimate for the density parameter Ω is then more or less independent of the value of H_0 . The best estimates of the density parameter obtained from dynamical estimates of the mass of spiral galaxy halos is

$$\Omega_{halo} \simeq 0.1. \quad (13.12)$$

Measuring mass on large scales

The techniques we have described provide estimates for individual galaxies. But what if there is substantial dark matter between the galaxies? Fortunately, we can apply similar principles to larger aggregations of matter, such as clusters of galaxies. Most galaxy clusters are almost certainly gravitationally bound; that is, their members orbit one another. For such a cluster, we can apply the virial theorem. There are still pitfalls, however. The virial theorem applies only to systems that are well approximated statistically; for a given cluster, it is not always easy to determine whether this state holds. Moreover, measurements of Doppler shifts provide only the radial velocity components, that is, the motion along our line of sight, not the full three-dimensional velocities of the galaxies; this introduces an additional uncertainty which must be taken into account. With all these caveats, and others, the results so far obtained indicate that

$$\Omega_{gc} \simeq 0.1 \text{ to } 0.3. \quad (13.13)$$

Another method of estimating mass densities is based upon the study of large galaxy clusters that show overall infall toward their centers. The nearest such large cluster to the Milky Way is the Virgo Cluster. The Virgo Cluster is the dominant mass aggregation in our immediate neighborhood and many smaller clusters, including our own Local Group, are falling toward it. It is possible to model this infall as a deviation from the general Hubble flow, and thereby to estimate the mass of the Virgo cluster. Such observations indicate

$$\Omega_{Virgo} \simeq 0.1 \text{ to } 0.2. \quad (13.14)$$

Recent work has hinted that there may be an even larger mass concentration somewhere beyond Virgo, in the direction of the constellation

Hydra, toward which Virgo and all its entourage are in turn falling. If this so-called “Great Attractor” is real, the models of infall toward the Virgo Cluster are probably incorrect, as the attractor would distort the simple motions expected; this might render invalid the current estimates of the mass-to-light ratio in the Virgo Cluster. Large-scale galactic motion remains a very active area of research.

Going to larger and larger scales, we find ourselves on increasingly shaky ground to use dynamical methods. Various preliminary measurements based on techniques such as galaxy surveys, the cosmic virial theorem, and so forth, generally obtain a density parameter well below unity. While these results must yet be regarded as tentative, it seems safe to say that none of the dynamical estimates indicate $\Omega_M = 1$.

It is at least possible to obtain a truly cosmological estimate of the total *baryonic* mass density. Chapter 12 discusses the theory that utilizes the fact that the rate of production of deuterium and other light elements in the early universe is very sensitive to the density of potential reactants. Only nucleons (protons and neutrons) take part in these reactions; hence measurements of primordial nuclide abundances, such as deuterium, imply a density due to nucleons. Since nucleons make up the overwhelming majority of baryons, it can be stated with some confidence that the density parameter of baryons is

$$\Omega_b \simeq 0.019 h^{-2}, \quad (13.15)$$

where $h \equiv H_0/100$ is the Hubble constant in units of $100 \text{ km s}^{-1} \text{ Mpc}^{-1}$.

Summarizing all the dynamical and nuclide evidence forces us to the inescapable conclusion that most of the matter of the universe not only is invisible to us, but is not composed of what we consider ordinary matter. From the gravitational dynamics of galaxies and galaxy clusters, we find that $\Omega_M = 0.2\text{--}0.3$. Note that this range of values overlaps nicely with both the supernova and the CBR data represented in Figure 13.8. The estimates from primordial nucleosynthesis indicate a much smaller number; using the current best value for H_0 this value is approximately $\Omega_b = 0.04$. Apparently ordinary matter, the stuff of atoms, makes up only 10–20% of the dynamical mass content of the universe. The rest must consist of some other type of matter, some massive subatomic particle that is not a baryon. This dark matter could fall into two categories; if it has low average energy and tends to clump together into self-gravitating balls on the scale of galaxies, it is termed **cold dark matter** (CDM). On the other hand, if the dark matter has high energy and thus moves through the universe at high speeds, it resists the tendency to clump. This more evenly distributed dark matter is called **hot dark matter** (HDM). The nonbaryonic dark matter content of the universe, Ω_{DM} , may well be a combination of both types. We shall discuss the nature of this mysterious dark matter in greater detail in Chapter 15.

Baryon density from big bang nucleosynthesis

Most of the gravitating mass in the universe is not in the form of baryons

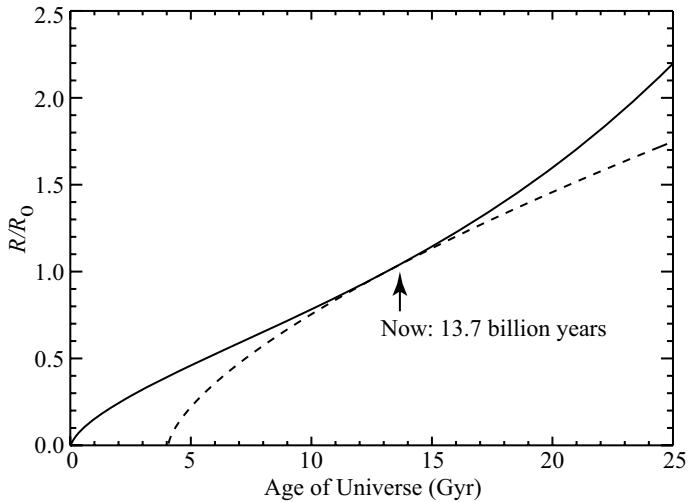


Fig. 13.10 Plot of R versus time for a model using the cosmological parameters from the concordance model. The dashed line shows the evolution of a flat Einstein-de Sitter universe for comparison. The cosmological constant increases the age of the universe compared to the standard model.

Lambda or dark energy

Cosmologists have known for decades that the universe was *nearly* flat, that it was extremely isotropic, and thus by inference it was homogeneous. These conditions seem very special; how was the universe initialized in such a unique way? Standard cosmology offers no explanation. However, in 1981 Alan Guth proposed the first version of the *inflationary* cosmological model.⁸ The theory of inflation proposes that the universe underwent a brief period of extreme exponential expansion very shortly after the beginning of time. This expansion smooths out any space curvature, and after it ends the universe is left with a flat geometry. Thus the inflationary model predicts $\Omega = 1$ naturally. Observers, however, were unable to find evidence for sufficient matter to be consistent with this value of Ω . The nucleosynthesis model seemed most definitely to rule out a dense universe composed of baryons. Dynamical measurements of mass indicated that a mysterious dark matter must be prevalent, and astronomers were left with the uncomfortable proposition that not only is the universe not dominated by ordinary atoms, but the nature of its major constituent is unknown. Adding to the overall discomfort was the age problem. The value of the Hubble constant was creeping up from its previously accepted lower value, reducing the Hubble time and creating a conflict with globular cluster ages. The best model seemed to be a low-density open model, even if the theorists were not fond of it. But as the errors in the measurements were being reduced, the values of the cosmological parameters converged; it began to seem possible that no standard model would be acceptable. One escape from this dilemma was to add a cosmological constant to the models, but cosmologists were reluctant to do so until there was better observational evidence in sup-

The inflationary model predicts a flat universe

⁸Chapter 16 treats this theory in detail.

port of such a phenomenon. The Type Ia supernova data, along with measurements of the tiny temperature variations in the CBR, have now provided that observational support. It appears that not only is there a cosmological constant, but its influence on the current universe is larger than that of the ordinary matter and the dark matter combined. Yet even though it is becoming apparent that Λ is a dominant aspect of the present and future universe, little is currently known about it.

Einstein's theory of general relativity is encompassed in his master equation, which states that the geometry of space-time is determined by its matter and energy content. When Einstein attempted to prevent his static cosmological model from contracting due to its own gravity, he noticed that it was possible to add a term to the equation that would, in effect, create a repulsive force over long distances, but would not affect properties such as conservation of mass-energy and consistency with Newtonian gravity. This term was the cosmological constant. For Einstein it was a mathematical constant that indicates how gravity behaves as distance goes to infinity. Under the assumption that gravity drops to zero at infinity, a seemingly sensible notion, it follows that $\Lambda = 0$. But there are other possibilities, and in Einstein's time there were no data on gravity's effects over huge distances. In fact, there are two interpretations of the cosmological constant's role in the Einstein equations of general relativity, one more geometrical and the other more physical. If Λ is placed on the left-hand side of the equation, as Einstein did, it enters as a geometry term. However, it could equally well be placed on the right-hand side, with the terms that account for energy, pressure, and matter. When grouped with the stress–energy terms on the right, it represents some sort of universal energy density; specifically, it would be an energy density of space itself, a **vacuum energy** density, which we shall write as ρ_Λ . This interpretation arises from a quantum-mechanical viewpoint.

If ρ_Λ is not equal to zero, then the larger the volume of empty space, the greater the energy. This means that it is energetically favorable to increase the volume of space. Hence a vacuum energy density acts like a *negative* pressure. (With ordinary positive pressure, the energy is increased by compressing a volume, much as a gas becomes hot as it is squeezed.) The relationship between a density and its associated pressure is the *equation of state*; cosmologists prefer to express the equation of state in the form

$$P = w\rho. \quad (13.16)$$

Cosmological components we have considered so far include ordinary matter, which exerts no significant pressure ($w = 0$); photons, which do have a pressure, with $w = 1/3$; and now vacuum energy, which contributes a negative pressure $w = -1$. Positive pressure causes the expansion of the universe to decelerate. Conversely, negative pressure causes the expansion to accelerate.

If Λ is indeed a consequence of a vacuum energy, is there a theory that tells us what its value should be? We have no such theory at the present time. The only nonzero prediction that has emerged so far from quantum

Different interpretations of Λ

Λ as a vacuum energy

w , the equation of state parameter for dark energy

Parameters of the universe

Parameter	Symbol	Value
Hubble constant	H_0	72 km s ⁻¹ Mpc ⁻¹
Geometry	Ω_k	0 (flat)
Mass density	Ω_M	0.3
Baryon density	Ω_b	0.04
Dark Matter density	Ω_{DM}	0.26
Cosmological constant	Ω_Λ	0.7
Deceleration	q_0	-0.55
Age	t_0	13.7 billion years

Quintessence, an alternative to Λ

mechanics gives a value that is wildly incorrect.⁹ Since we have as yet no good theory for a vacuum energy, perhaps we should be careful about jumping to conclusions about the nature of Λ . Some cosmologists have proposed that it is not really a vacuum energy, but is some other substance that exerts a negative pressure. This substance has been dubbed **quintessence** after Aristotle's heavenly element. A quintessence exerts a negative pressure, but in the corresponding equation of state, w need not be equal to -1. A different value of w would mean that the density and the pressure of the quintessence would not be constant in time, but would change as the universe expands. Quintessence-dominated models can be constructed with different values of w and their predictions computed, in order to determine whether those models match the data better than those with a more conventional, fixed Λ .

However, at the present time we must admit that the source of Λ is unknown, and some cosmologists simply refer to it as the **dark energy**. It is certainly one of the most surprising cosmological results in recent times, and it is of tremendous interest to theoretical physics.

The era of precision cosmology

A major goal of cosmology has always been to provide the best possible overall description of the structure and history of the universe. New observations with highly sophisticated telescopes are bringing us within reach of this goal. In this chapter we have described several methods for measuring important cosmological parameters by various independent means. Although these certainly do not represent the final results, the table of the parameters of the universe lists a set of values that are consistent with the accumulating experimental evidence.

The uncertainties in all these measurements are still larger than are preferable. However, because the cosmological parameters, such as H ,

⁹Chapter 16 discusses this topic.

Ω_M , Ω_Λ , Ω_k , and the age of the universe, are so intertwined, a measurement of one places restrictions on the acceptable values of others. What is striking about the current results is the agreement among different experiments. Even with some relatively large error bars on the specific values, the accumulated evidence is now converging toward a particular type of model, the **concordance model**. Independent and reliable measurements of the Hubble constant, the mass density, the deceleration parameter, and the geometry, along with limits on the age of the universe, point toward a flat, accelerating universe dominated by a cosmological constant. What we call matter, the stuff of which humanity, the stars, and the planets are composed, makes up no more than a few percent of the dynamical content of the universe.

The capabilities of both ground- and space-based observatories are improving so rapidly that there is real hope that we will have these answers in the near future. Astronomers now speak tentatively of the “era of precision cosmology,” a time when we measure with increasing accuracy many of the fundamental parameters of the universe. As the dark energy shows, however, those measurements will not necessarily mean that we immediately *understand* everything we see in the universe. But that understanding, when it comes, will be built on the foundation of these precision measurements.

The concordance model represents the current best values for the cosmic parameters

Chapter Summary

Current models of the universe are based upon general relativity and the cosmological principle. Many models are possible within this framework. Each model is described by a set of potentially observable values that account for the universe’s structure and evolution. The Hubble constant is one such parameter; it describes the current rate of expansion. The density of matter ρ determines the strength of gravity, and how rapidly the expansion slows. The cosmological constant term, Λ , sets the amount of acceleration that the universe might be experiencing. The curvature k describes the geometry of the universe; that is, flat, spherical, or hyperbolic. These values are conveniently written in terms of their fractions of a critical value, that is, as Ω_M , Ω_Λ , and Ω_k . By an appropriate definition of the critical values, the Friedmann equation shows that the sum of these Ω values must be unity.

Each of these parameters is measurable, and recent developments have led to increasingly tight constraints on their values. Since the values of Ω_M , Ω_Λ , and Ω_k are not independent, the measurement of any two establishes the third. However, it is best to try to determine each value independently. Measurements of distances to high-

redshift galaxies provide a way to observe directly the change in the scale factor $R(t)$, as well as any universal acceleration or deceleration. Recent studies of Type Ia supernovae, which exemplify a particularly bright standard candle, have indicated a remarkable result: the universe is accelerating. This surprising result helps to explain an observation that had become increasingly problematic in recent years: the age of the oldest stars seemed to be greater than the Hubble time inferred from the best value of H_0 . This apparent contradiction is reconciled in an accelerating universe, however, because in that case the true age of the universe can be greater than the Hubble time.

The geometry of the universe can be measured by observing the angular size of a known length as a function of redshift. This is difficult to do because of the lack of independent knowledge of the proper sizes of objects at large redshift. However, fluctuations in the cosmic background radiation provide a known length at a redshift of 1000. The apparent angular sizes of these fluctuations have been measured by the *WMAP* satellite, and the results indicate that the universe is flat.

A direct measurement of the matter content of the universe is formidable because only a small fraction of matter is in the form of visible objects such as stars. However, the presence of mass can be deduced from the effects of gravity. For example, the rotational velocities of the disk in a spiral galaxy provide a way to measure its mass. These studies indicate that galaxies have large mass-to-light ratios; specifically, for each solar luminosity there are 10 to 30 solar masses. The conclusion follows that most of the mass in the universe is *dark matter*. Only a small fraction

of this is ordinary matter. The composition of the rest of the dark matter remains a mystery.

Perhaps the greatest current puzzle is the nature of the Λ term. Is it a true cosmological constant, such as was originally proposed by Einstein, or is it some new form of energy that evolves with time? At present we have no theory to tell us what Λ is or what its value should be. The increasing precision of our cosmological observations is now outpacing our theoretical understanding of the physics of matter and energy.

Key Term Definitions

parameters of the universe A set of measurable quantities that describe and distinguish the full set of homogeneous and isotropic models.

dark matter Matter that is invisible because it emits little or no light. Most generally, dark matter includes both ordinary baryonic matter and any exotic forms of matter. Most of the mass of the universe is dark.

emission distance The distance to the source of light at the time the light was emitted.

reception distance The distance to the source of light at the time the light was received.

redshift-distance relation A theoretical relationship between the redshift of an object, such as a galaxy, and its distance from us. By measuring both distances and redshifts it is possible in principle to determine the evolution of the scale factor, $R(t)$.

standard candle An object of known intrinsic luminosity, useful in the measurement of luminosity distances.

angular size The angle subtended by an object on the sky. For example, the angular size of the full Moon is about 30 arcminutes.

kinematical method A method of measuring the mass density of the universe indirectly, by means of overall parameters of the universe such as its expansion rate. Kinematic methods exploit the fact that expansion rate, deceleration parameter, density, and curvature are not completely independent quantities but are related by the Friedmann equations, possibly extended to include a cosmological constant.

mass-to-light ratio The ratio of the total mass of a luminous aggregate of matter expressed in solar masses to its total luminosity expressed in solar luminosities.

dynamical method A method of measuring the mass of a galaxy, cluster, or even the universe, which makes use of the gravitational interactions of two or more bodies.

dark halo A massive aggregation of nonluminous matter of unknown kind that surrounds and envelopes galaxies.

virial theorem A statistical result that relates the mean gravitational field of a cluster to the dispersion of the velocities of the members of the cluster.

cold dark matter A form of nonbaryonic dark matter that has low energy and low particle velocities at the time it decouples from other matter early in the history of the universe. Such matter tends to clump gravitationally into galaxy-sized structures initially.

hot dark matter A form of nonbaryonic dark matter that has high energy and high particle velocities at the time it decouples from other matter early in the history of the universe. Such matter tends to clump gravitationally into large galaxy-cluster-sized structures initially.

vacuum energy The energy associated with empty space, that is, the vacuum itself.

quintessence A hypothetical exotic form of matter or energy that produces a negative pressure and a cosmological acceleration like a cosmological constant. A quintessence need not be constant in time, so the effective Λ force can change as the universe evolves.

dark energy The term given to the energy that is responsible for the overall acceleration of the universe. Possible dark energies include a cosmological constant Λ , a nonzero vacuum energy, or otherwise unknown forms of energy dubbed *quintessence*.

concordance model A model of the universe that shows the best overall agreement with data from a variety of observations, including redshift-distance tests, CBR fluctuations, and big bang nucleosynthesis calculations.

Review Questions

- (13.1) What is the age of an Einstein-de Sitter model whose Hubble time is 12 billion years? If the oldest stars are found to be 14 billion years old, what is the maximum possible value for the Hubble constant in an Einstein-de Sitter universe?
- (13.2) Using Figure 13.8, estimate the values of Ω_Λ and Ω_M for a flat universe with a Hubble time of 15 billion years and stars that are 12 billion years old. Is this universe accelerating or decelerating now?
- (13.3) What is the difference between *reception distance* and *emission distance*? What kind of information would you need, in addition to redshift, in order to determine these quantities exactly for objects such as quasars?
- (13.4) Consider the universe at a redshift of $z = 2$. If two galaxies were separated by a distance ℓ at the time corresponding to this redshift, what is their separation today?
- (13.5) Explain how the measured angular size of galaxies as a function of redshift could be used, in principle, to determine the geometry of the universe.
- (13.6) Where is most of the mass in spiral galaxies located? From what evidence do we draw such a conclusion?
- (13.7) Assume that the Sun is located 8 kpc from the Galaxy's center, and that it orbits at a velocity of 220 km s^{-1} . Using equation (13.8), estimate the Galaxy's mass in solar masses interior to the Sun's orbit. How much larger is the mass computed at twice the radius with the same orbital velocity?
- (13.8) Describe two distinct approaches to measuring the mass density of the present universe. How do these results compare with the accepted density due to baryons (ordinary matter)? To what conclusion does this lead?
- (13.9) What are two interpretations of Λ ? What is *quintessence* and how does it differ from the conventional Λ ?

This page intentionally left blank

A Message from the Big Bang

14

Key Terms:

- Olbers' paradox
- cosmic background radiation (CBR)
- recombination
- surface of last scattering
- seed perturbations
- Sachs–Wolfe effect
- first acoustic peak
- photon damping
- concordance model
- reionization

[T]hey looked for dung but found gold, which is just opposite of the experience of most of us.

Ivan Kaminov, on Penzias and Wilson's discovery.

Darkness and the expanding universe

When the Sun sets the night sky darkens, lit up only here and there by points of lights we now know to be distant suns. But *why* is the sky dark at night? The sky is bright during the day because the Sun is so close, and therefore its light fills the sky when it is visible; at night, conversely, it might be argued that the distant stars cannot brighten the sky. But this argument is inadequate, a fact which Kepler was one of the first to recognize. If the universe is infinite, and contains an infinite number of stars that live forever, then every line of sight must end on a star. It is true that a star's light diminishes as the square of the distance, but the volume of space sampled increases by exactly the same factor as distance becomes greater. Thus the night sky should be everywhere as bright as the average surface of a star; both night and day would be ablaze. Yet we do not observe this, and this has important cosmological implications.

Kepler was certain that this paradox required that the universe be finite. This resolution was satisfactory to him, but the explanation later ran aground in the Newtonian universe. The Newtonian model required balancing gravitational attractions equally in all directions, in order that the universe not collapse. Newton therefore assumed the existence of an infinite space, filled uniformly with an infinite number of stars. He believed, or perhaps hoped, that this arrangement solved the problem of gravitational collapse; about this he was quite wrong. Not only is the infinite Newtonian universe gravitationally unstable, but it also exacerbated the problem of the night sky. Edmund Halley tried to banish the paradox by attributing darkness to the remoteness of the majority of the stars, but this argument fails. Even if a particular star may be invisible to the eye due to its great distance, its light would combine with

The mystery of darkness at night

the light from an infinite number of other imperceptible stars, together accumulating to light the sky. Were it not for the fact that nearer stars completely block the light from those behind them, the sky would be infinitely bright in Newton's universe.

Another explanation was proposed as early as 1744 when Jean-Phillipe Loys de Chesaux attributed the darkness of the night sky to absorption by a fluid he imagined to permeate all space. Nearly 80 years later, Heinrich Olbers repeated this argument. Even though he was not the first to try to explain the darkness at night, for some reason the name of Olbers stuck to this awkward difficulty, and it has become generally known as **Olbers' paradox**. However, John Herschel showed in the middle of the 19th century that Loys de Chesaux' and Olbers' explanation was incorrect. Any fluid that filled the universe and absorbed starlight would, according to the laws of thermodynamics, heat up until its temperature was equal to the average temperature of the stars; it would then radiate just as much light as if it were itself a source of starlight. Herschel himself favored a hierarchical view of the universe, in which stars clump into clusters, the clusters bunch into larger clusters, and so forth *ad infinitum*. In a hierarchical universe, or for that matter in any nonuniform distribution of stars, there do exist lines of sight that are empty; this is the salient feature that distinguishes uniform from nonuniform. But if the universe is to be isotropic and homogeneous on the large scales, as the modern view requires, then a strictly hierarchical model is ruled out.

Olbers' paradox hung over cosmology well into the 20th century. With the discovery of the expanding universe, many cosmologists immediately accepted the cosmic expansion as the answer. The light from the most distant stars is so redshifted that it contributes no appreciable energy to lighting our skies. Some authors have gone so far as to assert that the darkness of night is sufficient proof that the universe is expanding. However, Edward R. Harrison has emphasized that the resolution of the paradox does not require expanding space. The crucial flaw in the traditional argument was the assumption that the stars could shine forever. With our modern understanding of energy conservation, we know that this could not possibly be the case. Light carries energy, and thus stars must liberate energy in order to shine. Stellar lifetimes are very finite. When we look sufficiently far into space, and therefore back into time, eventually we look to an era before any stars existed. Moreover, in any universe which is not infinitely old, or which expands, the size of the *observable* universe is finite, because of the finite speed of light. The finite volume of the observable universe contains a finite number of stars, so most lines of sight never intersect the surface of a star at all. Even if multiple generations of stars live and die, the sky will still be dark. The number of stars is too small, and stellar lifetimes are simply too short, to fill the vastness of space with light. The darkness of the night sky quite elegantly rules out the simple model of an infinite universe filled with infinitely old stars.

Stars do not shine forever

However, more light than that which originates with the stars permeates the universe. The cosmic background radiation fills the sky in all directions, yet its wavelengths lie below the range that is visible to human eyes. The expanding universe *does* tell us something about this modern version of Olbers' paradox. As is the case for the light from distant quasars, the expansion of the universe has caused the cosmic background light to redshift and lose energy on its long journey across the universe. But if the light has now redshifted to lower energy, must there not have been a time that its energy was high? Could the universe have once been filled with hot photons of visible light? If so, the cosmos was once ablaze throughout but has become dark due to the cosmological redshift. The fact that we cannot see the cosmic background radiation demonstrates that the universe is expanding. Thus the darkness of the night sky, a simple fact of life that most of us have taken for granted since childhood, is seen to yield an important clue to the structure of our universe. In cosmology, the most innocuous phenomena can sometimes prove to be very profound.

The universe is filled with radiation

Noise from the sky

In 1964, Arno Penzias and Robert Wilson of Bell Laboratories were searching for the source of some weak noise observed in the signal detected by a sensitive radio antenna in Holmdel, New Jersey. The antenna had originally been built for communications via the satellite *Echo*, but Penzias and Wilson, who were radio astronomers, planned to study radio emission from the Galaxy. In order to map such an extended source, it was necessary for them to characterize all potential causes of noise in their receiver, so as to be able to subtract that noise from the desired signal. They began their calibration with a wavelength much shorter than the radio wavelengths anticipated to originate from Galactic sources, expecting that any noise in the microwave band would be due to their receiver, or to the Earth's atmosphere. Accordingly, they chose a wavelength of 7.35 cm for their initial tests. Penzias and Wilson felt confident that this would enable them to evaluate any noise due to the antenna's electrical circuits, or to radiation from the atmosphere. Much to their surprise, however, a persistent excess noise remained after they had accounted for every source they could identify. In their determination to find the origin of this noise, they went so far as to dismantle part of the receiver in the spring of 1965, cleaning it thoroughly and removing the residue from a pair of pigeons that had been nesting in it. Yet despite their best efforts over many months, the excess noise remained.

Radio astronomers describe their signals by, roughly speaking, fitting the radiation to a blackbody spectrum¹ regardless of whether the original source emits blackbody radiation or not; the use of such an *antenna temperature* enables the signals to be standardized, and provides

A radio signal can be characterized by a temperature

¹See Chapter 4 for a description of the properties of blackbody radiation.



Fig. 14.1 Penzias and Wilson with the horn antenna with which they discovered the cosmic background radiation. (Lucent Technologies' Bell Laboratories, courtesy AIP Emilio Segrè Visual Archives, Physics Today Collection.)

a reference for comparison purposes. The enigmatic noise discovered by Penzias and Wilson corresponded to an antenna temperature of approximately 3.5 K, a relatively small signal but still larger than they had anticipated for electrical noise. If the atmosphere had been responsible, the amplitude of the noise would have varied from the zenith to the horizon in a predictable way. Atmospheric emissions coming from near the horizon would travel through a much thicker layer of atmosphere than would those originating directly overhead.² But the background noise was found to be independent of the direction in which the antenna was pointed, ruling out an atmospheric cause. Neither did it vary with the time of day or year, which was evidence against an origin in the Galaxy, or in any other anisotropic celestial source. Given that the noise could not be attributed to the antenna circuits themselves, such constancy in space and time indicated a *cosmic* origin, but in 1965 few scientists anticipated that the cosmos itself might hum with microwave energy.

At nearly the same time and only a few miles from Holmdel, P. James E. Peebles, a young theorist at Princeton University, had just carried out a calculation that predicted low-temperature radiation from the early universe. He and his colleagues Robert Dicke, P. G. Roll, and D. T. Wilkinson were even in the process of constructing a receiver specifically to look for this background radiation when word of the work of Penzias and Wilson arrived in Princeton. Dicke, the original designer of the Holmdel receiver, had already arrived at the idea that relic radiation might be present from an early phase of the universe. However, he had based his expectations not upon a hot big bang, but upon a cyclic model

²A similar effect accounts for the reddening of the Sun at sunrise and sunset; light emitted near the horizon traverses more atmosphere and thus has more blue photons scattered out of it.

in which the universe is expanding from an earlier state of collapse; he was looking for evidence of element *destruction*, not creation. Yet Peebles' eventual theoretical work amounted to a rediscovery of Gamow, Alpher, and Herman's much earlier work on nucleosynthesis,³ with more realistic assumptions. In any case, the Princeton group was on the right track and would surely have discovered the background radiation had it not been for the serendipitous, but timely, results of Penzias and Wilson. Thus the interpretation of these mysterious emissions as the overall **cosmic background radiation (CBR)** arrived promptly. The discovery of the CBR was the most significant cosmological observation since Hubble's results, and earned for Penzias and Wilson the 1978 Nobel Prize in physics.

Given the importance of the CBR in establishing the standard big bang model, it is interesting that a determined search for it was not carried out much earlier. Although Gamow, Alpher, and Herman's contributions to big-bang nucleosynthesis had been widely recognized, their prediction of a background of low-temperature radiation throughout the universe was not. Ya. B. Zel'dovich improved upon Gamow's results in the early 1960s, about the same time that Peebles was performing his calculations independently. A. G. Doroshkevich and I. D. Novikov even suggested in 1964 that microwave radiation might be sought to check Gamow's theory. However, the Russians never pursued the matter further. Better communication among theorists and observers, as well as among different groups of scientists, might have speeded up the discovery of the cosmic background radiation somewhat. Regardless, it is clear that the scientific foundations were in place by the time that Penzias and Wilson announced their results.

In retrospect, it was realized that the CBR had actually been detected indirectly as early as 1941. Interstellar gas clouds often contain molecules as well as atoms; and molecules possess discrete energy levels, just as do atoms. In general, the spectra of molecules are much more complex than those of atoms, since not only can electrons jump around, but the molecule as a whole can rotate and/or vibrate; nevertheless, molecules can also be identified uniquely by their spectra. In 1941, W. S. Adams observed transitions of cyanogen (CN) in a molecular cloud between the Earth and the star Zeta Ophiuchus. From these data, A. McKellar found that one line in the spectrum of the cyanogen could be explained only if the molecules were being excited by photons with an equivalent temperature of approximately 2.3 K. At the time, no explanation could be found for this phenomenon, so it simply disappeared into the sea of scientific information. Only in 1965 did George Field, I. S. Shklovsky, and N. J. Woolf realize the significance of this observation. In 1993, K. C. Roth, D. M. Meyer, and I. Hawkins again took spectra of cyanogen in several clouds between the Earth and nearby bright stars in an intentional search for the CBR excitation, finding a

The discovery of the CBR

³See Chapter 12.

Two great cosmological discoveries of the 20th century

temperature of 2.729 K, in excellent agreement with other measurements.

The two greatest cosmological observations of the 20th century were the discovery of the expansion of space and the discovery of the cosmic background radiation. In both cases, the discoveries were astonishing and revolutionary; but with hindsight it is clear that cosmological theory was prepared for them. Einstein's theory of general relativity, and the difficulties in obtaining a static model, provided an immediate theoretical interpretation for Hubble's finding. Similarly, the existence of the cosmic background radiation was anticipated by cosmologists investigating the early history of expanding models. The explanation of the background radiation as a relic from a hot, dense phase in the history of the universe was sufficiently persuasive to create a coalescence in cosmological theory; indeed, it is a primary piece of the evidence that makes the big bang model the standard.

Traveling photons

The universe becomes transparent

The cosmic background radiation is the most direct data we have from the early era of the universe's existence. It stringently constrains the permissible models for the formation of the universe and its constituent structures. It tells us immediately that the universe was once very much hotter than it is today, which by itself is convincing evidence for the hot big bang. The hot big bang model proposes that early on, the matter density and temperature were very high and the universe was opaque; matter and radiation constantly exchanged energy. Thus the universe was in thermal equilibrium in its early stages of existence and would naturally have been filled with blackbody radiation appropriate to its temperature. As the universe expanded, cooled, and rarefied, there came a time at which the electrons were captured by atomic nuclei. Free electrons easily scatter photons, while electrons bound to atoms do not. At **recombination** the atomic matter ceased to interact strongly with the radiation, and the universe became transparent to light. At this point the state of thermal equilibrium between the photons and the matter ceased, and the blackbody radiation streamed freely through space. Since then the universe has been transparent to photons, and the cosmic radiation has traveled unimpeded through space. Occasionally a few of its photons strike receivers located on a small planet in the Milky Way Galaxy. The background radiation has been redshifted by the universal expansion, just like photons from distant galaxies, so the CBR photons now have energies much lower than they had upon their emission. This would not affect the *shape* of the spectrum, however. All blackbody spectra have exactly the same shape, differing only by their amplitudes and their peak wavelengths, both of which depend only upon the temperature of the radiation. Therefore, any process that affects all photons in the spectrum equally, as does the universal expansion, cannot change the shape of a blackbody spectrum but can only shift it. Since

the CBR photons have all been equally redshifted to low energies, we still see the spectrum as blackbody today, but one corresponding to a very low temperature,

Except for their very rare interactions with matter, these cosmic photons are neither created nor destroyed; they simply stream through space in all directions. Consider some typical region of the universe; by the cosmological principle, this volume must be completely representative of the universe as a whole, as long as it is sufficiently large for true isotropy to prevail. On average, there will be some number of photons per unit volume; this is the *photon number density*, which we shall denote by n . We can characterize each photon's energy by specifying its wavelength λ and using the equation $E = hc/\lambda$, where h is Planck's constant. Our representative region will thus contain some amount of energy due to the photons, which we would obtain by summing the energy of each of the photons in this volume. But h and c are constants for all space and time; hence we can always write the total energy per unit volume, that is, the energy density, as $\mathcal{E} = nhc/\lambda_a$, where λ_a is the representative wavelength of all the photons in the volume. (More precisely, it is the wavelength corresponding to the average frequency of the spectrum.) Expansion changes the energy density in two ways. First, the wavelength of any individual photon is redshifted by the overall expansion. Recall that $\lambda \sim R$ for any wavelength; the representative wavelength λ_a redshifts in exactly the same way, since R is the same for all photons at any cosmic time. This allows the factor of $1/\lambda_a$ in the energy density to be expressed in terms of the scale factor. The second effect of expansion is that the fixed number of photons occupies a larger and larger space as the volume increases; thus the photons become more and more diluted. Therefore, the number density n of photons decreases due to the increase in volume from the expansion. Since the volume increases as R^3 , we have $n \sim 1/R^3$, just as for matter density. Combining these two effects, we find that the energy density in the photons of the cosmic radiation decreases as the *fourth power* of the scale factor, that is, $\mathcal{E} \sim 1/R^4$.

We can carry this simple calculation further. The energy density of blackbody radiation is also proportional to the fourth power of the temperature, $\mathcal{E} \propto T^4$. Substituting our previous formula for the behavior of the energy density as a function of the scale factor, we obtain the remarkably simple result that the temperature of the cosmic radiation diminishes as the inverse of the scale factor,

$$\frac{T_{\text{then}}}{T_0} = \frac{R_0}{R_{\text{then}}} = (1 + z). \quad (14.1)$$

This equation shows that the temperature of the CBR was higher at earlier and earlier times, corresponding to greater and greater redshifts.

It is possible to measure the temperature of the universe at high redshift. In 1994, a group of astronomers used the Keck telescope atop Mauna Kea in Hawaii in a search for excitations in the atoms of clouds at high redshift. The approach is exactly like that carried out in the early

CBR photon wavelengths are redshifted

CBR photon density is diluted by expansion

The CBR temperature-redshift relation

Observations of distant molecular clouds test the temperature-redshift relation

studies of cyanogen in nearby clouds that found unexplained molecular excitations, later realized to be due to the background radiation; however, the high-redshift study used a transition in atomic carbon that was more appropriate for this measurement. The group obtained spectra from two clouds lying close to a distant quasar, at a redshift of $z = 1.776$. The measurement was difficult, since the clouds were remote and spectra demand the collection of considerable light. Fortunately the Keck, with its 10-meter primary mirror, is capable of collecting a sufficiently large quantity of photons from such a distant source. The astronomers used a very long exposure and an extremely high-resolution spectrograph. Their results were gratifying; the excitation of one cloud corresponded to a temperature of $10.4 \text{ K} \pm 0.5 \text{ K}$, while that of the other indicated a temperature of $7.4 \text{ K} \pm 0.8 \text{ K}$. From equation (14.1), the predicted temperature of the background radiation at the redshift employed is 7.6 K. Much of the difference between the measured and the theoretical temperatures is most likely due to the interference of molecular collisions and other phenomena within the clouds themselves, which complicates the interpretation of the data. This method has subsequently been extended to clouds even farther away, and the results are again consistent with theoretical predictions. By themselves, these measurements, or any other single set of measurements, cannot prove the big bang. These experiments simply provide more support by confirming a prediction of the theory.

Where has the energy gone?

The CBR loses energy with time

If the cosmic background radiation has redshifted to lower temperatures throughout most of the history of the universe, with a corresponding decrease in the energy of each photon, where has that energy gone? In fact, what happened to the energy lost by *any* redshifted photon traveling through the universe? Is it not the case that energy, taken in all its forms, is conserved?

The behavior of the thermal photons from the big bang is similar to that of the particles of an expanding gas. When a gas expands, it cools; similarly, as the universe expands, the temperature of the photons is reduced. Although this line of reasoning seems perfectly sound, it fails when we try to extend it to account for the lost energy. When an ordinary container holding a gas, which might be a gas of photons, expands, then something must cause it to do so. If the gas particles themselves cause the expansion, such as in the cylinders of an internal-combustion engine, then some of their internal (heat) energy must be converted into the work required to expand the container. If an external agent, such as a motor pulling on a piston, causes the expansion, then similarly some work is exerted, so that overall energy is still conserved. However, there is no such external agent in the cosmos, nor is there a boundary against which the photons push. The photons themselves are certainly not driving the expansion of the universe. We have learned

from general relativity that any positive pressure, including that due to photons, would contribute to gravity and thus would tend to make the universe *collapse*, not expand. In any case, the background radiation is nearly perfectly isotropic, and thus there are no pressure changes from one point to another to create any photon push.

What about the gravitational field itself? In the case of a photon climbing from a gravitational field around a massive object, we could understand the redshift qualitatively by imagining that the photon was consuming its intrinsic energy to gain gravitational potential energy, much as a tossed ball rising in a gravitational field loses kinetic energy. Unfortunately, this conceptualization does not work for cosmological models. The gravitational potential energy is a consequence of the change in a gravitational field in *space*. The universe as a whole is spatially homogeneous and isotropic at all times in its existence, so there is no spatial change in the field; indeed, in the case of a flat universe there is not even a spatial curvature, but there is still an expansion redshift.

Does this mean that the universe violates the conservation of energy? The principle of conservation of energy that is familiar to physicists is a *local* statement, known to hold only for finite regions. Cosmological energy is, by definition, quite nonlocal. A major impediment to our understanding is that we currently do not even have a consistent definition of total cosmological energy. We cannot formulate a conservation law for a quantity we cannot define. It may not even be possible in principle to define such a cosmological energy, in which case there is no reason to expect that any corresponding law of energy conservation will exist. If this is true, we need not be concerned with the lost energy of the redshifted photons. On the other hand, this rather glaring exception to an extraordinarily fruitful scientific principle may tell us something. We know that general relativity is an incomplete theory because, as it stands, gravity cannot be melded with the other three fundamental forces of nature. Those three forces all do conserve energy. In the theories that explain the other three forces, the conservation of energy arises because their laws and equations are indifferent to whether time runs forward or backward; these theories are *symmetric* under time reversal. In contrast, the universe does not appear to be time symmetric; there is an arrow of time in the evolution of the cosmos, running from the low-entropy big bang to the high-entropy heat death of the future. If time symmetry is required for energy conservation to hold, then the universe as a whole simply may *not* conserve energy. Perhaps if we achieve the “theory of everything” that unites all four fundamental forces, we will find that the grandest laws of physics are not symmetric under time reversal. We may then be able to formulate some more complete conservation law, or else we shall understand why this is not possible. Unless and until we reach this summit, the missing energy of the redshifted photons must remain unexplained.

The total energy of the universe is undefined

Studying the cosmic background

The CBR is a medium for studying the big bang itself

What is the importance of the cosmic background radiation to cosmology? Most obviously, the cosmic background radiation provides evidence for the big bang itself. This alone qualifies it for the title of a “great discovery.” Moreover, the CBR is among the very few phenomena that can tell us about the conditions in the very early universe. When we observe the background photons, we are looking back to a time approximately a million years after $t = 0$, when the (re)combination of electrons with nuclei allowed photons to stream freely. Since the time of recombination the CBR photons’ energy has redshifted, with the peak wavelength dropping from the visible portion of the spectrum, where it was located at recombination, to the microwave region. The present temperature of the background radiation is close to 3 K. Recombination occurred when the temperature was around 3000 K; from these facts, the redshift formula leads to the conclusion that recombination took place approximately at $z = 1000$. This redshift represents a fundamental limit on our ability to look into the past with telescopes; we will never be able see directly through the impenetrable state of the universe that existed prior to this time. The last instant that the universe was opaque is effectively the edge of the visible universe for us. This edge is called the **surface of last scattering**. Particle reactions were over within a few seconds after the big bang; the hot, opaque, matter–photon plasma then dominated the universe for an interval on the order of a million years. Since we can make no direct observations of this interval, we can only hope to understand what happened before recombination, and particularly in the first few seconds, by studying the imprint of the events that occurred then upon the universe we observe now.

The CBR supports the cosmological principle

The cosmic background radiation is also the best evidence we have that the universe adheres to the cosmological principle. If the universe is truly homogeneous and isotropic, the relic cosmic background radiation should be a perfect blackbody in all directions, excluding any possible interactions with matter lying between its distant source and our radio antennas. But is the spectrum of the CBR really consistent with blackbody radiation? A blackbody spectrum is produced by a dense gas in perfect thermal equilibrium; that is, the energetics of the gas is fully described by a single temperature. The specification of that temperature determines precisely what should be observed at every wavelength in the spectrum. Penzias and Wilson observed the background radiation at only a single wavelength. Although the energy measured at that wavelength was appropriate to a blackbody of around 3 K, they did not actually know whether the radiation they discovered was truly part of a blackbody spectrum or not. The shape of the spectrum could be determined only by taking data at many wavelengths.

However, we know that the universe is not *perfectly* homogeneous on all scales today; galaxies and clusters of galaxies obviously contain a higher density of matter than exists in intergalactic space. Today’s galaxies and galaxy clusters had their origins in slightly overdense re-

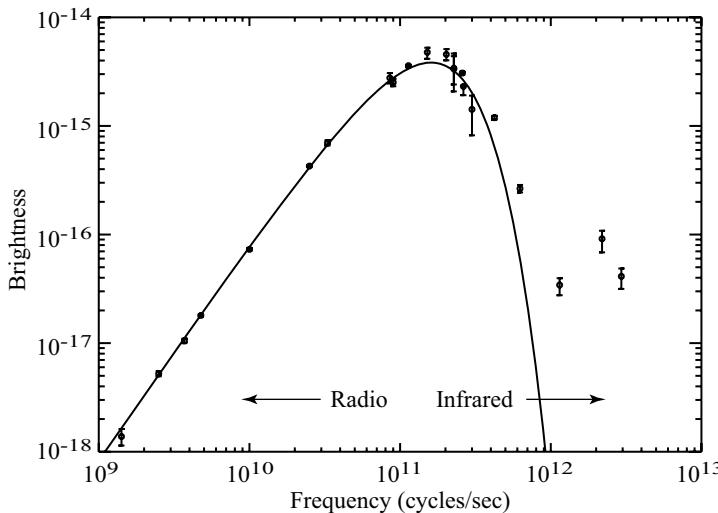


Fig. 14.2 Observations of the brightness of the CBR (in units of $\text{ergs s}^{-1} \text{cm}^{-2}$ per unit of solid angle) at various frequencies, obtained before the launch of the *COBE* satellite. The radio observations fit a blackbody with a temperature of 2.7 K, but the Earth's atmosphere is opaque in the infrared region in which the spectrum should turn over. Observations in this part of the spectrum were taken by balloon- or rocket-borne telescopes. In the 1980s these observations suggested excess infrared radiation might be present.

gions of gas in the early universe. Therefore, although the universe must have been highly homogeneous at the time the CBR was emitted, it cannot have been completely homogeneous even then. Any such irregularities in the gas produced slight variations in the temperature of the CBR, and since there has been essentially no further interaction of the CBR photons with matter, those temperature fluctuations have been preserved to the present day. This implies that the CBR carries important information about the structure of the universe; thus it is important to discern how isotropic the background radiation actually is. After Penzias and Wilson's discovery, astronomers faced two important questions: is the spectrum of the CBR truly a blackbody, and does the spectrum exhibit any anisotropies, that is, variations in temperature across the sky?

Scientists took up the challenge and began to measure the intensity of the CBR over a full range of wavelengths, and in different directions on the sky. Unhappily for the astronomers, however, this is not an easy task from a location beneath the blanket of the Earth's atmosphere. The atmosphere is very nearly opaque to several regions in the electromagnetic spectrum. To confirm a true blackbody spectrum, it was particularly important to measure the CBR into the infrared, at frequencies above the radio band, since the peak emission of a 3 K blackbody lies in this region. Unfortunately, in this part of the infrared spectrum water molecules in the atmosphere block almost all the radiation; the only means of observing this region of the spectrum is to go above the Earth's atmosphere. For nearly 25 years, observers flew receivers to the top of the atmosphere on balloons and rockets, searching for the elusive and important higher-frequency emissions. Such experiments were exceedingly difficult and complex, and were subject to numerous systematic and instrumental errors. Figure 14.2 shows the state of the data from this pioneering research, as of the late 1980s. The infrared

Is the CBR a true blackbody?

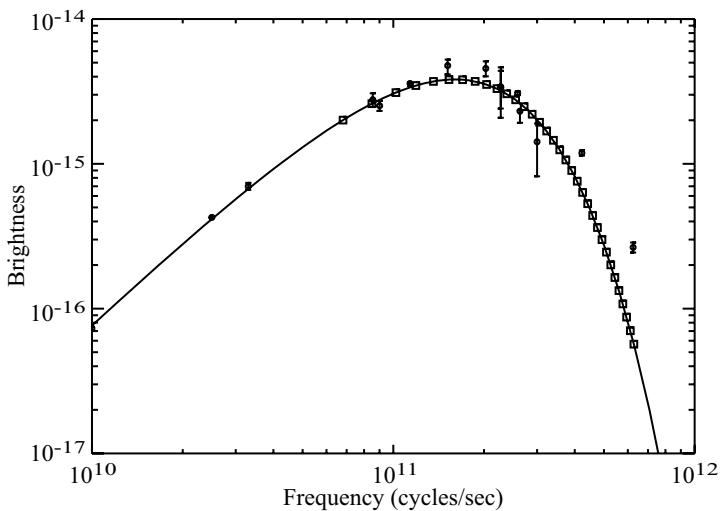


Fig. 14.3 Observations of the brightness of the CBR (in units of $\text{ergs s}^{-1} \text{cm}^{-2}$ per unit of solid angle), measured over the critical range of infrared frequencies by the *COBE* satellite. The *COBE* data are the squares; they are fitted by a blackbody curve with a temperature of 2.725 K. The observations show no evidence for any deviation from a perfect blackbody; in fact, the squares used here are substantially too large to represent the true error bars. The pre-*COBE* data are included for comparison.

measurements seemed to suggest possible deviations from a blackbody, although the radio data indicated that the background is well fitted by a blackbody.

Clearly, the best hope for measuring the CBR with high accuracy was to place a receiver aboard a satellite, far above the atmosphere and its obfuscating effects. This view was provided by the *Cosmic Background Explorer (COBE)* satellite, which was launched in 1989. *COBE* was able to measure the intensity of the CBR across a broad range of infrared wavelengths without interference from the atmosphere. The results were spectacular. The cosmic background radiation was found to obey a perfect blackbody law to better than 0.03%, an impossible precision before the satellite observations became available. The temperature of the CBR was at last confirmed to be 2.725 K, with an uncertainty of ± 0.002 K.

The fidelity of the CBR to a blackbody spectrum is a powerful vindication that the radiation originates from the universe itself and not, as some rival theories had proposed, from a general background of stellar and gaseous emissions. It would be impossible to obtain such a perfect blackbody spectrum over the whole sky from a combination of dim sources, none of which would be expected to be at the same temperature as any of the rest nor even, by itself, exactly a blackbody. The sum of many discrete sources could be a blackbody only if the emitted photons were brought somehow into equilibrium, or *thermalized*, by various interactions. However, models that invoke such effects are contrived and artificial.

The other important question is the amount of *anisotropy* in the CBR, that is, how much, if any, does its temperature vary in different directions. This is also a difficult measurement to make from the ground. In addition to absorbing much of the electromagnetic radiation impinging upon the Earth, the atmosphere varies in density and other properties,

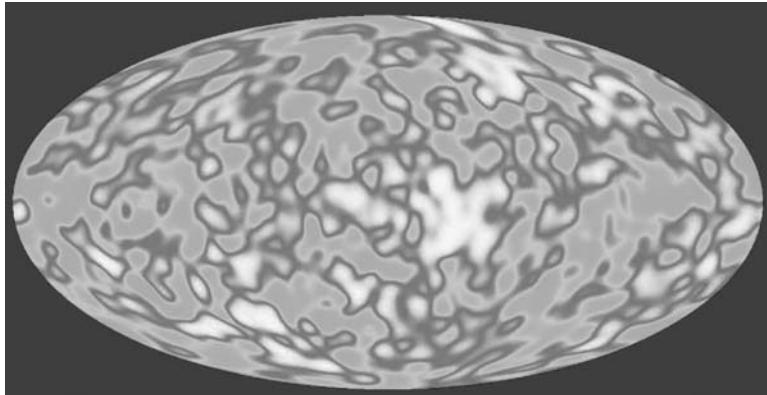


Fig. 14.4 *COBE* map of the sky showing temperature variations in the CBR. The temperature variations are very small, less than 20 millionths of a kelvin. (Goddard Space Flight Center/NASA.)

complicating comparisons from one direction to another. Balloon and rocket data indicated that the CBR had the same temperature, within the experimental errors, at all points in the sky. Such measurements were valuable but were, obviously, very prone to errors. Careful comparisons of the temperature of the sky at different directions in the radio bands had found no variations down to about one part in 10^5 , but only better data from a satellite could settle the question. *COBE* included a special instrument, called the Differential Microwave Radiometer, that simultaneously compared the radiation coming from two directions at three different frequencies. This device mapped the full sky and found that the temperature of the cosmic background radiation was nearly the same in all directions, to a level of precision unattainable in earlier experiments. This provides the best confirmation available that on the largest scale, the universe is very isotropic. We may then conclude from the Copernican principle that it is also homogeneous. The observed isotropy of the CBR puts stringent limitations upon any cosmological model we might construct; no matter what, it must always accommodate this fact. This evidence justifies our assumption that the Robertson–Walker metric provides a good first approximation to the universe, and we can continue to work out the details within the context of models based upon this metric.

COBE also found the limit to this isotropy. More detailed statistical analyses of the data indicated the presence of very small anisotropies, near the level of sensitivity of the instrument. The temperature fluctuations are quite small; the measured variation in temperature from one part of the sky to another is less than about one part in 10^5 . The effective resolution of *COBE* was too poor to see fluctuations on scales smaller than about 7° on the sky. However, *COBE* established that these variations exist; the detailed study of the fluctuations on small scales is the goal of subsequent CBR-related space missions.

The launch of COBE answered the question of whether the CBR spectrum was a true blackbody

The universe is isotropic to a high degree

COBE discovered fluctuations in the CBR temperature

Where are we going?

The dipole CBR anisotropy traces our motion through the cosmos

The raw data from *COBE* do show some significant nonisotropic effects that are well understood. For example, the plane of the Milky Way itself stood out in the *COBE* data; dust within the galaxy emits considerable infrared radiation. Nevertheless, it was simple to subtract away this effect, leaving only measurements of the cosmic background radiation. Another aspect of the observed CBR anisotropy that was easy to understand was the overall Doppler shift due to our motion through the cosmos. This anisotropy is said to be *dipole* because it has two well defined and opposite points: the point of largest blueshift indicates the direction in which we are heading relative to the CBR, while the point of greatest redshift is immediately opposite. Between the two extremes there is a smooth and systematic transition over the sphere of the sky. This is a well known, well understood phenomenon that is easily taken into account when searching for the more subtle anisotropies. These smaller anisotropies, which are indicative of tiny variations in matter density present in the early universe, are distributed over the entire sky. The nature of these small-scale fluctuations has much to tell us about conditions in the early universe. The dipole anisotropy is of less intrinsic interest, since its origin is not mysterious; however, the data from *COBE* enabled scientists to determine the Earth's motion through the universe with unprecedented accuracy.

This is a significant piece of data, since if we wish to study the motions of external galaxies we must first determine our own peculiar velocity. The *COBE* data showed that we are moving at about 390 km s^{-1} in a direction toward the region of the sky that is assigned to the constellation Leo. Many components contribute to this overall motion. The Earth orbits the Sun with an average speed of approximately 30 km s^{-1} . The Sun itself orbits the center of the galaxy, with a speed of about 220 km s^{-1} . The Milky Way, in turn, moves through space due to gravitational interactions with the other Local Group galaxies, as well as due to infall into the Virgo Cluster. Finally, the Virgo Cluster itself may have a systematic motion. The net sum of the shifts corresponding to all these motions yields the net Doppler shift relative to the cosmic background radiation.

Does this imply that the CBR constitutes a special frame, like the ether of the 19th century was thought to define? In a sense, the CBR *does* define a special frame: the frame that is at rest with respect to the overall matter distribution of the universe. If the Milky Way had no peculiar motion, but was simply being carried away from other galaxies solely by the Hubble flow, we would see no Doppler shift in the spectrum of the CBR. This standard of rest in turn provides a convenient definition of the cosmic time, such as appears in the Robertson–Walker metric. Yet the existence of cosmic time might seem itself to violate the equivalency of frames. Does this somehow repudiate special relativity?

The first response to this question is simply that we can define any convenient frame we wish, and we can always note that we are moving

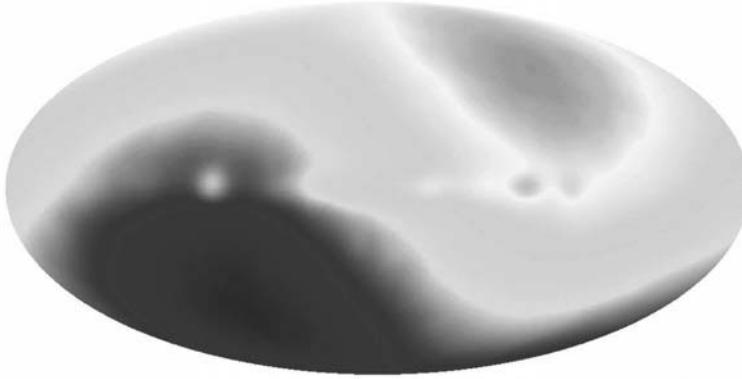


Fig. 14.5 *COBE* map of the sky showing the large-scale temperature variation in the CBR due to the motion of the solar system with respect to the cosmic background. The temperature difference corresponds to a velocity of approximately 390 km s^{-1} . (Goddard Space Flight Center/NASA.)

relative to such a frame. More significantly, the cosmic rest frame defined by the CBR results from general, not special, relativity, and we have learned that special relativity can be valid only in localized regions of space-time if gravitation is present, as it always is. The cosmic rest frame represents the inertial frame of observers who are freely falling in the large-scale gravitational field of the universe; that is, observers who are moving only with the Hubble flow. The specialness of the frame of the CBR should be a consequence of the overall distribution of mass in the universe, and does not conflict with the special theory of relativity.

The assertion that the frame of the CBR defines the cosmic rest frame is testable. We can ask whether the frame in which distant galaxies recede isotropically coincides with the frame of the CBR, after subtracting away the dipole anisotropy. If both phenomena are the result of an isotropic expansion, and if there are no intrinsic anisotropies in the CBR, then this should certainly be the case. There is no particularly good explanation for any possible skewing of the background radiation with respect to the Hubble flow, but a few suggestions have been advanced, so it is appropriate to check whether our fundamental assumptions will hold up. Unfortunately, the answer to this question is not so easily obtained. The *COBE* measurements determine the *net* motion of the solar system relative to the CBR but, as we have discussed, that net motion is composed of many subcomponents. We must somehow extricate the overall motion of the Milky Way, the other galaxies in the Local Group, and the motions of nearby clusters, such as the Virgo Cluster, all relative to the most distant galaxies. Only after we subtract the peculiar motion of the Milky Way can we determine the frame of the Hubble flow.

Like so many cosmological observations, this measurement is quite difficult to make. The Earth's motion is analogous to that of an amusement-park ride that simultaneously rotates about an axis and travels in some arbitrary direction, while nearby background objects move as well. All the galaxies in our immediate vicinity, and indeed for a distance of at least several megaparsecs, have their own, ill determined peculiar velocities that confound measurements of the Milky Way's motion. The project is further complicated by the fact that measured redshifts pro-

The cosmic rest frame

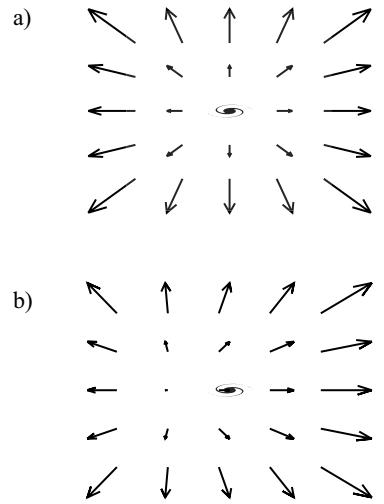


Fig. 14.6 Plotting the velocity of nearby galaxies surrounding the Milky Way as vectors illustrates the appearance of (a) simple Hubble flow and (b) Hubble flow distorted by the peculiar velocity of the Milky Way. The actual situation is more complicated, since every galaxy has its own peculiar motion.

vide only the radial component of the velocity relative to us. That is, we can determine whether a galaxy is approaching or receding along our line of sight, but we cannot observe its complete, three-dimensional motion. We must simply do our best to sort out all these motions.

The first step is to account for the well known revolution of the Earth around the Sun. Next we must consider the motion of the Sun in the Galaxy. Since the total mass of the Galaxy within the solar orbit is not independently known, we cannot apply Kepler's third law directly to compute this orbit. (The observations of solar movement are actually used for the opposite purpose: from the measured velocity, we compute a Galactic mass by means of Kepler's law.) To plot the Sun's orbit, astronomers must carefully analyze the velocities of globular clusters to find the center of the Galactic motion. After this center is established, the motions of stars near the Sun are analyzed; since these stars travel on orbits very close to that of the Sun, the Sun's motion about the Galactic center can then be determined. The result is that the Sun travels in the plane of the disk on a very nearly circular orbit about the Galactic center with a speed of approximately 220 km s^{-1} . Knowing the solar orbit enables us to correct for the motion of the Sun within the Galaxy, in order to pick out the inherent motion of the Milky Way.

The next step is to accumulate redshifts of galaxies that are, cosmically speaking, nearby. It is then possible to plot a map of vectors representing their radial velocities, with corrections for the local solar motion. In principle, scrutiny of such maps enables us to deduce the motion of the Milky Way, and even of the Local Group as a whole. It should not be surprising that it is difficult to carry out such a program; nearly every group of astronomers that has attempted it has obtained somewhat different results. A recent development gives an example of the importance of this task. Several analyses carried out since 1988 have indicated a bulk motion toward something called the "Great Attractor," a hypothetical mass concentration in the direction of the constellation Centaurus, beyond a large galaxy cluster called Hydra-Centaurus.

At some point all galaxy motions should blend into a background Hubble flow, but until we can determine the peculiar motion of our and other galaxies, it is not so easy to extract the pure Hubble flow. We can plot velocity-redshift diagrams for distant galaxies located in all directions, and attempt to shift the data until the scatter of the points is minimized; but so far this approach is not really definitive. For the present we shall simply state that the data are consistent with the coincidence of the CBR frame and the frame of the Hubble flow. Work in this area continues.

Mapping galactic motions

Ripples in space

The dipole anisotropy tells us our net motion through the cosmos. What do the smaller-scale temperature fluctuations tell us? Prior to the launch of *COBE*, astronomers had carried out extensive ground-based searches

for these anisotropies, but had found no definite detections. Yet if the universe is so fantastically homogeneous, how could any overdense structures, such as galaxies and galaxy clusters, have formed? Gravity is always attractive, which implies that any small overdensity in a uniform background will tend to be amplified. Initial **seed perturbations** in density are still required, however, since even in the presence of gravity, a perfectly uniform density distribution will remain uniform for all time. Furthermore, these initial fluctuations must have been large enough to produce galaxies at a very early stage. Galaxies have been found at redshifts of $z = 6$ and higher, which places important constraints on any theory of structure formation. Computing the lookback time for such remote galaxies shows that they existed when the universe was no more than about a billion years old. This implies that galaxy formation began promptly after recombination. Hence, regardless of the way in which clusters, superclusters, and the empty regions between them may have formed, their origins as fluctuations *must* have been present during the radiation-dominated era of the big bang. Cosmologists therefore greeted with relief the news that *COBE* had detected temperature fluctuations.

Although the amplitudes of these perturbations are small, their traces are visible in the cosmic background radiation we observe today. The anisotropies that we see in the cosmic background radiation ultimately had their origin in the earliest moments of the universe. How is it that we can see the ghosts of primordial structures in the present background radiation? If the CBR radiation consists only of free photons, by what mechanisms might the primordial matter perturbations have affected it? How do these fluctuations tell us something about the properties of the universe? To answer these questions, we must return to the earliest moments of the universe, this time not to follow the average evolution but to track the evolution of small deviations from the average. During the radiation era, the universe was filled with hot plasma. The photons constantly collided with matter particles, keeping radiation and matter tightly coupled; together the radiation and matter behaved like a single hot, dense gas. In such a plasma, gravitational forces compete with radiation pressure, but those forces only cause accelerations if there are differences in density or pressure. For example, if a region of plasma is cool and slightly overdense, gravity tends to pull it together; as the region contracts the plasma heats. The additional pressure produced by the heating pushes outward, halting the contraction and reversing it. The resulting expansion cools the gas, reducing the pressure and allowing the cycle to repeat. This oscillating battle between gravity and pressure is basically a sound wave, albeit in a somewhat unfamiliar form. Sound waves, like other waves, are characterized by the scale of an oscillation, the *wavelength*, the period of an oscillation, the *frequency*, and the strength of the oscillation, the *amplitude*. A full spectrum of such waves at all wavelengths and frequencies is present during the radiation era, but the behavior and amplitude of each of these waves depend on the physical conditions of the early universe and the parameters that govern the cosmic evolution.

Sound waves in the sky

Gravitational perturbations give rise to forces that generate oscillating sound waves, but at very early times such waves were frozen in place; they could not oscillate due to the finite speed of sound and the finite age of the universe. The sound speed in a hot relativistic plasma is about 60% of the speed of light. When the universe was very young, even waves with such a high speed had not yet had sufficient time to move an appreciable distance. The wavelength of a sound wave is approximately equal to the distance that can be traveled at the speed of sound during one oscillation cycle. Thus in the early universe, the longest wavelength that could act was set by the maximum distance that a sound wave could have traveled over the age of the universe. Initially these were very short wavelengths, but as time passed ever longer waves were unfrozen and began to move.

Gravitational forces are produced by regions of enhanced dark matter and baryon density, but in the early universe those matter densities were much less than the energy density of the radiation, and the forces due to radiation pressure greatly exceeded those due to gravity. Only after approximately 10^{12} seconds, when the radiation era gave way to the matter era, was the matter able to begin to clump under its own gravity. The radiation was still coupled to the matter, but at this point gravitational forces started to overwhelm the forces due to pressure, and matter began to control the evolution of the universe. The matter perturbations were then able to grow in amplitude, but their growth was limited; once the photons responsible for the pressure were released from the matter to stream freely through space, the sound waves ceased oscillating. This occurred at the time of recombination, at approximately 10^{13} seconds, when the free electrons combined with nuclei and the universe became transparent. Although only about 300,000 years separated the beginning of the matter era from the recombination epoch, this was the critical interval during which the universe stamped its character on the sound waves that moved through it. When we observe the CBR from the time of recombination, we see the pattern of those waves imprinted on the sky.

The major process by which this imprinting occurs is the **Sachs–Wolfe effect**, which is essentially gravitational redshifting and blueshifting. Prior to recombination, the photons and the matter interacted constantly. After recombination, the photons streamed unimpeded through space, carrying the memory of their last scattering. Photons that last scattered from a region of higher than average density were forced to climb out of the slightly stronger gravitational well and thus are redshifted, relative to the average, while photons that last scattered from a lower-density region are blueshifted. This alone would tell us very little, except for the fact that these redshifts and blueshifts are due to a factor distinct from the cosmic expansion and thus do not affect all photons equally. Consequently, the Sachs–Wolfe shifting appears to our measuring devices as a difference in temperature of the scattered photons, in comparison to the smooth background. This causes the temperature of the CBR to show deviations from the perfectly uniform, isotropic back-

CBR temperature fluctuations reveal perturbations in the early universe

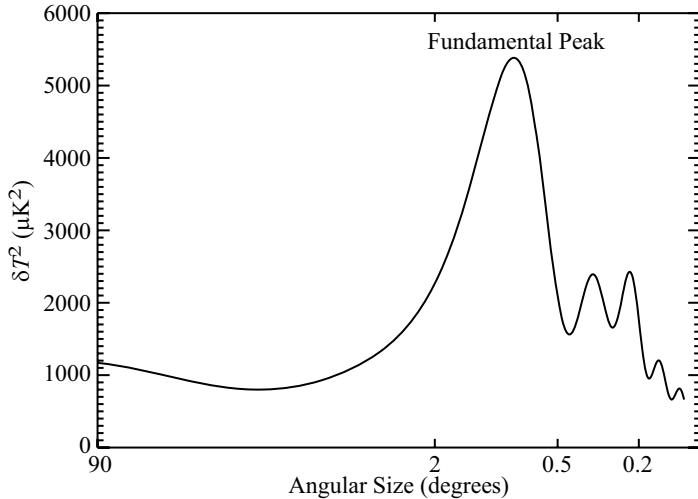


Fig. 14.7 Power in the CBR temperature fluctuations (in squared microkelvins, μK^2) versus angular size on the sky, as predicted by the current best model. The fundamental peak is located at approximately 1 degree. At smaller angular scales are the first and second harmonic peaks. Beyond them are additional peaks whose amplitudes are increasingly damped.

ground temperature. Relative to the background density, the sizes of the primordial density fluctuations are roughly proportional to the amplitude of these temperature anisotropies; that is, $\delta\rho \propto \delta T$, where δ is the symbol indicating the perturbation in the variable that follows it.

The temperature variations in the CBR appear to be noise; they reflect the noise of the cosmic sound waves. If all the notes that combine into noise are of equal amplitude, the result is said to be *white noise*. Studies of the CBR anisotropies have shown that they do not represent white noise; certain wavelengths have greater amplitudes. Different models of the early universe make specific predictions about the evolution of the perturbations, which leads to different predictions of the amplitudes of the temperature fluctuations δT at different angular scales on the sky. Comparison of the predictions with the observations provides a means of testing models describing a time in the evolution of the universe that we cannot directly observe.

Figure 14.7 summarizes the result from a particular calculation of the temperature fluctuations corresponding to the current best model of the universe. The graph shows the fluctuation amplitudes as a function of angular size projected onto the sky. The larger the angular size, the longer the wavelengths. The figure exhibits considerable complexity, but parts of it are fairly straightforward to understand. The maximum distance that a sound wave could have traveled in the 300,000 years from the beginning of the matter era to the time of recombination is called the *sound horizon*. A wavelength comparable to this horizon distance will correspond to an oscillation period of 300,000 years. This particular wave will have completed exactly one cycle when we observe it. Of greatest interest, however, is the wavelength of the wave that completed *half* an oscillation cycle over this interval. Such a wave had just enough time to reach its maximum compression before recombination halts its oscillations. Thus this wave represents the hottest (and coldest) temperature

Long sound waves have not yet completed one oscillation

anisotropies seen in the CBR. The peak in power corresponding to this wave is the *fundamental peak*, and its angular size, approximately 1 degree on the sky, is called the *fundamental scale*. To the left of this peak are the wavelengths that are too long to have completed even half an oscillation prior to recombination. The wave amplitudes corresponding to those long wavelengths will simply reflect the initial perturbations, more or less unchanged from the earliest moments that they existed. To the right of the fundamental peak are those waves of shorter wavelength and higher frequency that also were in a state of maximum compression at recombination. The wavelengths of these waves will be odd-integer fractions of the fundamental scale. Between these peaks are those corresponding to wavelengths that are even-integer fractions of the fundamental scale; these waves were in a state of maximum rarefaction at recombination.

Those waves that were in a state of either maximum compression or maximum rarefaction at recombination produce the most significant fluctuations on the CBR. The smaller the angular scale on the sky that these waves subtend, the more oscillations the waves underwent before recombination, and the more they could be affected by various physical processes operating in the early universe. These scales can tell us a great deal about the properties of the universe. The fundamental scale, that is, the loudest note, is called the **first acoustic peak**. It corresponds to a particular wavelength, λ_f . This peak is located at a specific angular size projected onto the sky; in Figure 14.7 its scale is around 1 degree. The angular size-distance relationship connects λ_f to this scale; in its simplest form this is given by $\theta = \lambda_f/d$, where d is the distance to the surface of last scattering. In Chapter 13 it was shown that the angular size we observe is a function both of distance *and* of the spatial geometry of the universe (see Figure 13.6). The wavelength of the fundamental scale can be computed fairly simply from the physics of the early universe. The corresponding angular size that we observe in the CBR depends sensitively on the geometry of the universe. This fundamental scale provides the standard length that was missing from previous applications of the angular size-redshift test. It follows that determining the angular size of the first acoustic peak in the CBR yields a direct measurement of the curvature of space.

The fluctuations in the CBR tell us more than the spatial geometry of the universe. The fundamental scale depends upon the sound speed, which itself is a function of the matter density. Thus an increase or decrease in the matter density would shift the peaks by means of its influence on the sound speed, with a faster sound speed resulting in a shift to the left and a slower sound speed shifting the peaks toward the right. Since the densities are related to the expansion rate of the universe, the acoustic peak locations are functions of the Hubble constant. The amplitudes of the sound waves also carry information about the structure of the universe. These amplitudes are partially determined by the opposing forces of gravity and radiation pressure, which depend sensitively on the densities of baryons, dark matter, and radiation. In a universe

The fundamental scale provides a test of geometry

with a greater baryon density, gravity will be stronger and the maximum compression will be greater. This will be reflected in the amplitude of observed peaks that are in compression, such as the peak at the fundamental scale. Higher baryon density will also tend to increase the amplitude of the compression peaks over the amplitude of the rarefaction peaks. Another important phenomenon is *photon diffusion*. Before recombination, photons could diffuse only slowly through the matter. The distribution of the photons thus tended to be more uniform and homogeneous than the distribution of the matter alone. Hence the tight coupling between photons and baryons meant that photons were able to squelch perturbations in the matter, an effect known as **photon damping**. This effect reduces the amplitudes of the shorter-wavelength sound waves, those whose oscillation times are comparable to the period over which recombination takes place. The rapid decrease in the amplitudes of the perturbations that is visible at the extreme right-hand side of Figure 14.7 is a consequence of photon damping.

In practice, cosmologists must consider a wide range of physical effects when predicting the fluctuations in the CBR; the present discussion has provided only a suggestion of the many interrelationships among the various constituents of the universe. The aim of modeling the fluctuations is to test the theoretical predictions against the observations. The *COBE* experiment demonstrated that these perturbations exist. Subsequently, a number of careful studies were carried out using a variety of ground-based and balloon-borne detectors. These experiments provided the first evidence for the existence of the first acoustic peak and for some of the higher-frequency oscillations. But the definitive data came from *COBE*'s spacefaring successor, the satellite *WMAP*.

While matter and radiation were tightly coupled, photon diffusion tended to squelch density perturbations

The results from WMAP

The *Wilkinson Microwave Anisotropy Probe*⁴ was launched on June 30, 2001. After taking up its station at a point 1.5 million kilometers from Earth, the satellite began to measure the tiny temperature differences in the CBR between two particular points in the sky. Over an interval of time the satellite scans the entire sky to produce a complete map of the CBR temperature fluctuations. An example is shown in Figure 14.8.

WMAP was designed to measure temperature differences as small as $0.2 \mu\text{K}$, with an angular resolution of about 12 arcminutes. This high spatial resolution sets it apart from *COBE*, which had a resolution of approximately 7 degrees. The primary scientific data produced by *WMAP* are a set of maps of the CBR in several different microwave frequency bands. After correcting for a number of effects, such as the presence of the Milky Way, the data are analyzed to determine which angular sizes

⁴The satellite was renamed in honor of David Wilkinson following his death in 2002. Wilkinson was one of the pioneers in the study of the cosmic background radiation at Princeton in the early 1960s, and he remained a leader in the development of CBR experiments throughout his career.

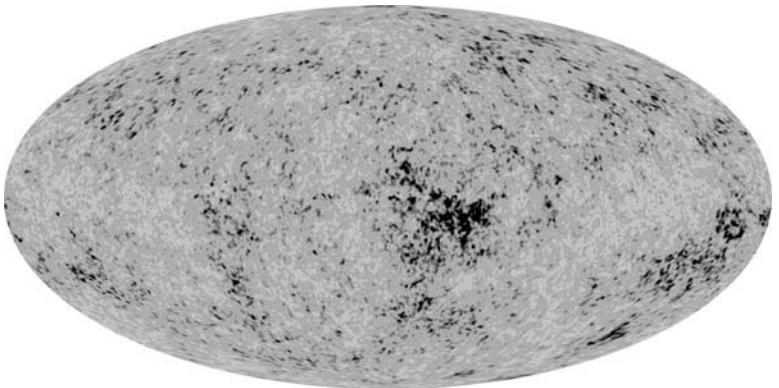


Fig. 14.8 A map of the sky showing the distribution of temperature fluctuations in the CBR, as measured by the *WMAP* satellite. The sizes of the temperature fluctuations depicted range over $\pm 200 \mu\text{K}$. (NASA/WMAP Science Team.)

contain the largest fluctuations. This *angular power spectrum*, depicted in Figure 14.9, can then be compared with theoretical predictions to determine the best-fit model of the universe. In constructing this figure, the *WMAP* data are supplemented by higher-resolution measurements obtained from the ground by the *CBI* and *ACBAR* microwave telescopes.

One of the immediate results from these data was the improved measurement of the angular size of the first acoustic peak. This observation has the advantage of sidestepping many of the uncertainties inherent in obtaining galactic distances; it measures the geometry of space directly. The scale of the first acoustic peak can be imagined to form one edge of an enormous isosceles triangle on the sky, with the light paths from the two ends of the peak to Earth forming the other two sides. The shape of this triangle is determined by the geometry of space lying between the Earth and the surface of last scattering. Using the *WMAP* data, the angular size-distance relationship indicates that the geometry of the universe is flat to within 2%. The flatness of the universe poses a puzzle, since dynamical methods of measuring mass density have long indicated that Ω_M is well below one. However, the results from measurements of distant Type Ia supernovae had already provided evidence that the universe is *accelerating*, implying a nonzero Λ . This dark energy can be responsible for a flat geometry in a low-matter universe.

In computing a model to compare to the CBR data, either we can work solely with the *WMAP* power spectrum, or else we can bring all of our cosmological data to bear. It is possible to construct a range of significantly different models that would fit the *WMAP* data, but some of these models would be quite inconsistent with other cosmological observations. Many other sources of solid cosmological data exist, including the improved measurements of the Hubble constant from *HST*, the limits on baryon density from big bang nucleosynthesis calculations, the known limits on stellar ages, the dark-matter density, and the supernova distance measurements. By combining all these data we can create a **concordance model**, our current best measurement of the

The concordance model

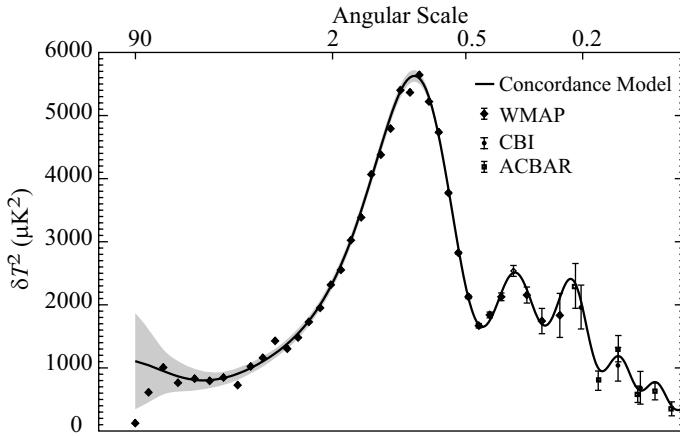


Fig. 14.9 Angular power spectrum of temperature fluctuations as measured by *WMAP*, supplemented by ground-based observations from the *CBI* and *ACBAR* microwave telescopes. The curve illustrates the predictions of the best-fit concordance model. The gray shading shows the range of variation expected from observing the universe at a random location. (Adapted from a plot by the NASA/*WMAP* Science Team.)

parameters of the universe.⁵ This model states that the universe is flat ($k = 0$), began 13.7 billion years ago, contains a significant cosmological constant (or dark energy) component, given by $\Omega_\Lambda = 0.7$, and has a mean mass density of $\Omega_M = 0.3$, of which only 13% is ordinary baryons. It is remarkable how well all the data fit together, even though they derive from quite distinct considerations and experimental methods. For example, the baryon density parameter is measured both by the abundance of deuterium and by its effect on the relative amplitudes in the second and third acoustic peaks; yet there is close agreement between these measurements.

The acoustic-peak data from *WMAP* helped determine the cosmic parameters to unprecedented accuracy, but the observations were more or less in line with earlier results obtained from ground-based and balloon-borne measurements of the CBR fluctuations. However, *WMAP* also provided a few surprises. Among these was the first determination of the time of **reionization**. At recombination, electrons combined with nuclei to form hydrogen atoms, but in the present universe most of that intergalactic hydrogen is ionized. It follows that at some point much of the neutral hydrogen must have been reionized. This event is tied to the first appearance of stars in the universe, because the ultraviolet light from stars is the source of the energy that ionizes the hydrogen. Reionization can be detected in the CBR because some of its photons scattered off the freed electrons, and evidence of that scattering can be discovered in the polarization of the CBR. The data from *WMAP* indicate that reionization may have begun as early as 200 million years after recombination. This remarkably early star formation poses a challenge for theories of star and galaxy formation.

Another interesting feature of the data is the unexpectedly small value of the so-called quadrupole component, the data point at the far left-hand side of Figure 14.9. The best-fit model predicts a slight upturn at

The first stars

⁵See Chapter 13.

this end of the plot, yet the measured power drops off so much that the actual value lies well outside even the range expected from random variations within the universe (the gray band in the figure). This deviation had been seen in the *COBE* data as well, but at that time it was unclear whether it was significant, or whether it was within the expected range of variation or perhaps was due to some systematic error or foreground contamination. While such effects may yet be the explanation, it is also possible that this hints at some new physics not yet incorporated into our cosmological models.

We have every reason to expect that the precision with which we are able to determine the parameters of the universe will increase. The next major space mission planned to study the CBR is the *Planck* satellite, which is scheduled to be launched by the European Space Agency in 2007. The new satellite will have even greater sensitivity and higher angular resolution than has *WMAP*; it should continue the process of refining this cosmic story written on the sky.

Chapter Summary

One of the classic questions of historical cosmology is Olbers' paradox, which asks why the sky is dark at night. The finite age of the universe and the finite lifetimes of stars provide the answer, but it was discovered that the universe is filled with light in the form of the cosmic background radiation, or CBR. The cosmic background radiation was discovered in 1964 by Arno Penzias and Robert Wilson, who had originally intended to use a radio telescope to study Galactic emissions, but were puzzled by a persistent noise in the instrument that they could not explain. Once the CBR was recognized for what it was, scientists struggled for over two decades to measure its spectrum accurately. Did the CBR have a blackbody spectrum, as would be expected for an early universe in thermal equilibrium? Since much of the most interesting portion of the CBR spectrum is absorbed by the Earth's atmosphere, the final answer had to await measurements from space. The *COBE* satellite determined that the CBR does correspond to blackbody radiation, with a temperature of a little more than 2.7 K. The redshift formula for the CBR temperature specifies the variation of the blackbody temperature with redshift. For example, the blackbody temperature was 2700 K at $z = 1000$. This redshift corresponds approximately to the time when electrons combined with protons to form hydrogen atoms; as a result of this event, called recombination, the universe

became transparent, allowing the CBR photons to stream into space.

COBE also measured the Doppler shift due to the motion of the Earth and the Milky Way with respect to the CBR. After subtracting away this so-called dipole shift, *COBE* found evidence for tiny residual temperature fluctuations, which represent the imprints of the small inhomogeneities in the early universe that grew into the large-scale structure that we see today.

The *COBE* satellite could not observe the CBR with high enough spatial resolution to map out the temperature fluctuations in detail. This was accomplished by the more recent *WMAP* mission. *WMAP* was able to measure the level of temperature fluctuations at different angular sizes on the sky. These observations can be related to theoretical models of the growth of fluctuations in the early universe. The results have helped to define a *concordance model* of the universe, a model that provides the best fit to all the various pieces of cosmological data. The concordance model is a flat universe that is accelerating due to a dark energy component, $\Omega_\Lambda = 0.7$. The matter content of the universe, $\Omega_M = 0.3$, is dominated by an unknown nonbaryonic dark matter. Continuing observations of increasing accuracy hold the promise of obtaining cosmological parameters to high precision in the immediate future.

Key Term Definitions

Olbers' paradox The fact that the night sky is dark even though in an infinite universe with stars that live forever, the night sky would be as bright as the surface of a star. The paradox disappears when it is realized that stars do not live forever and the universe is not infinitely old.

cosmic background radiation (CBR) The black-body radiation, now mostly in the microwave band, which consists of relic photons left over from the very hot, early phase of the big bang.

recombination The moment in the early universe when the temperature became sufficiently low that free electrons could no longer overcome the electrostatic attraction of the hydrogen nuclei and were captured to form atomic hydrogen. When this occurred the universe became transparent.

surface of last scattering The point at recombination at which the CBR photons last interacted with the baryonic matter. After this, the CBR photons streamed freely through space. The surface of last scattering is what is seen when the CBR is observed.

seed perturbations The initial small fluctuations in the universe that grow to become the observed CBR temperature fluctuations, and eventually large-scale cosmic structure.

Sachs-Wolfe effect The scattering of photons from perturbations in the early universe. Photons that last interacted with an overdense region suffer a gravitational redshift, whereas those which last scattered from an underdense region are blueshifted.

first acoustic peak The longest wavelength maximum in the CBR fluctuations. This wavelength corresponds to a pressure wave in the early universe that has completed half an oscillation cycle, and hence has reached maximum compression, at the time of recombination.

photon damping The tendency of photons in the early universe to smooth out inhomogeneities in matter with which they are in thermal equilibrium.

concordance model A model of the universe that has the best overall agreement with data from a variety of observations, including redshift-distance tests, CBR fluctuations, and big bang nucleosynthesis calculations.

reionization The point in time early in the universe, but after recombination, when the first stars formed and their ultraviolet light began to ionize the neutral hydrogen gas that filled the universe.

Review Questions

- (14.1) What phenomenon accounts for most of the darkness of the sky at night? How does the expanding universe contribute to darkness at night?
- (14.2) Briefly describe how the cosmic background radiation was first discovered (with an antenna, not theoretically).
- (14.3) Why is the CBR considered to be the most definitive evidence available for the big bang?
- (14.4) If the CBR originated from the early moments of the universe, when conditions were much hotter than they are today, why is its temperature so cold in the present universe? What happened to the energy lost by the photons?
- (14.5) Explain the significance of the measurements of the *COBE* satellite. What is the shape of the CBR spectrum? Does it vary from one direction to another, in either shape or temperature? What anisotropies were found by *COBE*, both on large and small scales, and what is the significance of these anisotropies?
- (14.6) How do the observations of *WMAP* provide a direct measurement of the geometry of space?
- (14.7) The early appearance of galaxies subsequent to recombination is a challenge for theory. Which *WMAP* results suggest that galaxies formed very early in the history of the universe? The best estimate of the redshift at recombination is $z = 1090$. Assume the first stars appeared at a redshift of

$z = 17$. Galaxies have been observed at redshifts as large as $z = 10$. Using the following formula for the lookback time, in years, in the flat Einstein-de Sitter model,

$$t_{\text{lb}} = 13 \times 10^{10} \left(1 - \frac{1}{(1+z)^{3/2}} \right),$$

calculate the times corresponding to these three redshifts. How much time elapsed between the first stars and the observed galaxies?

Part V

The Continuing Quest

This page intentionally left blank

Dark Matter and Large-Scale Structure

15

Key Terms:

Through space the universe grasps me
and swallows me up like a speck;
through thought I grasp it.

Blaise Pascal

So far we have modeled the universe by treating it as mostly smooth and homogeneous, in keeping with the cosmological principle. In reality, the universe is not perfectly smooth; it contains stars, galaxies, clusters of galaxies, and even superclusters of galaxies. On the grandest scales, those appropriate to cosmological models, these local density enhancements can be averaged into a smooth background. Eventually, however, cosmologists must tackle the problem of the origin and evolution of structure in the universe.

The universe contains a great deal of structure, much of it far vaster than those small objects upon which we live or depend. From the cosmological perspective, what is wondrous is not the formation of stars and planets, which still occurs today, but rather the formation of larger aggregations of matter. Everywhere we look, we see galaxies. There are at least as many galaxies in the observable universe as there are stars in the Milky Way. Galaxies are the dominant luminous components of the universe. Telescopes are probing ever deeper into the cosmos, finding galaxies at greater and greater redshifts. Galaxies must have formed very early in the history of the universe, perhaps within a billion years after the big bang.

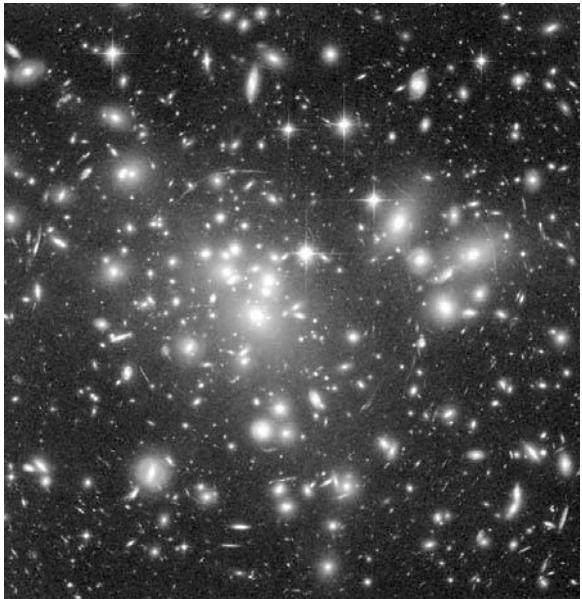
Galaxies, like humans, are more likely to be found in the company of others than alone. Many galaxies, including our own, dwell in a **galaxy cluster** of only a few to tens of members. The cluster to which the Milky Way belongs is called the *Local Group*. This group of galaxies, all of which are gravitationally bound and interacting, is dominated by two large spirals, the Milky Way and the Andromeda Galaxy.¹ Each of these two large spirals has several smaller satellite galaxies, which orbit their primary like moons orbit a planet. The Large and Small Magellanic Clouds, visible only from the Southern Hemisphere of the Earth, are two of the larger satellites of the Milky Way. The Local Group has a few other fairly significant members, such as a small galaxy in the

- **galaxy cluster**
- **supercluster**
- **top-down structure formation**
- **bottom-up structure formation**
- **dark age**
- **reionization**
- **galactic cannibalism**
- **dark matter**
- **MACHO**
- **neutrino**
- **hot dark matter**
- **WIMP**
- **cold dark matter**
- **Harrison–Zel'dovich spectrum**
- **photon damping**
- **collisionless damping**
- **correlation function**
- **void**
- **biased galaxy formation**

Galaxies are the dominant visible constituents of the universe

¹The Andromeda Galaxy is also known as M31.

Fig. 15.1 *Hubble Space Telescope* view of the massive galaxy cluster Abell 1689. The gravity of this cluster is so great that it acts as a gravitational lens, producing distorted images of galaxies located behind it at much greater distances. (Credit: NASA, N. Benitez (JHU), T. Broadhurst (The Hebrew University), H. Ford (JHU), M. Clampin (STScI), G. Hartig (STScI), G. Illingworth (UCO/Lick Observatory), the ACS Science Team and ESA.)



constellation Fornax and the modest spiral M33, as well as dozens of tiny dwarf galaxies. Many small clusters such as the Local Group exist in the universe.

The Local Group is a lonely outpost of only a few citizens, compared to the urban sprawl within some galaxy clusters. The nearest large cluster is the Virgo Cluster, 18 Mpc distant, with some 250 major galaxies and a few thousand small ones. The giant elliptical galaxy M87 occupies nearly the central spot of the Virgo Cluster. The Virgo Cluster is an example of an *irregular cluster*. Irregulars show no particular symmetry and consist mostly of spiral galaxies, with a few elliptical members. Often a giant elliptical galaxy resides at the center of the cluster. Beyond the Virgo Cluster, at a distance of over 50 Mpc from the Sun, is an unusually rich cluster in the constellation Coma Berenices. The Coma Cluster consists of several thousand large galaxies and an unknown number of smaller ones. It is an example of a *regular cluster*. Regular clusters are roughly spherical or ellipsoidal in overall shape and are dominated by elliptical galaxies, with few spirals. The correspondence between cluster type and typical galaxy type is intriguing, and may provide an important clue to the origin and evolution of galaxies and galaxy clusters.

The Local Group and several other small clusters are satellites of the Virgo Cluster. A cluster of galaxy clusters is called a **supercluster**. In keeping with the nomenclature, the supercluster consisting of the Virgo Cluster and its satellites, including the Local Group, is called the Local Supercluster. Many gravitationally bound superclusters are known; the organization of the universe seems to go from galaxy to cluster to supercluster. How did this hierarchy develop? Did superclusters form first, then fragment into clusters, which in turn shattered into individ-

The distribution of galaxies

How did galaxies and galaxy clusters form?

ual galaxies? This is called **top-down structure formation**. In an alternate scenario, **bottom-up structure formation**, galaxies are the fundamental building blocks; galaxies form, then become drawn into clusters by mutual gravitational attraction, followed by the evolution of superclusters as the clusters are themselves pulled together. There are other possibilities. Did mass concentrations the size of clusters form first, which then separated into individual galaxies while attracting one another into superclusters? A complete theory of structure formation should provide an explanation for what we see at all these scales, and should answer the question of which structures formed first.

The emergence of galaxies

All the structure seen in the galaxies, clusters, and superclusters has its origins in the earliest moments of the history of the universe. The hot, radiation dominated plasma that emerged from the big bang was remarkably smooth and homogeneous, but it nevertheless contained small density fluctuations, the seed perturbations of structure. At some point after the epoch of equal density, these perturbations began to grow and evolve in complicated ways. The traces of these primordial perturbations are now revealed in the pattern of small temperature fluctuations imprinted on the cosmic background radiation (CBR).² After recombination, matter and radiation went their separate ways. The matter perturbations continued to evolve, but there was no longer any light to delineate this evolution. The universe then entered a phase known as the **dark age**, which occurred prior to the time when the first stars formed and matter became luminous.

The origins of structure

The *WMAP* satellite's observations provided data that indicate when these first stars formed. Immediately after the recombination epoch the universe was filled with neutral gas, but when stars formed they became a significant source of ultraviolet light that *reionized* the gas. This **reionization** produced free electrons that could interact with and scatter from the CBR photons as they stream through space. *WMAP* can detect evidence of this scattering in the CBR. The satellite found that the first stars arrived on the scene at a remarkably high redshift of around $z = 17 \pm 4$, corresponding to a time only approximately 200 million years after the big bang.

Evidence for the first stars

Another line of evidence for the rapid formation of structure comes from quasi-stellar objects (quasars). Quasars are among the most ancient objects we can see. High abundances of iron, even greater than that in the solar neighborhood, have been found in the emissions of quasars at $z = 3.4$ and greater. Since iron originates only in stars and supernovae, this observation indicates that quasars must have formed in an environment in which stars had already existed long enough for some to have exploded as supernovae. Furthermore, there is little doubt that

²The evolution of the seed perturbations prior to recombination is described in Chapter 14.

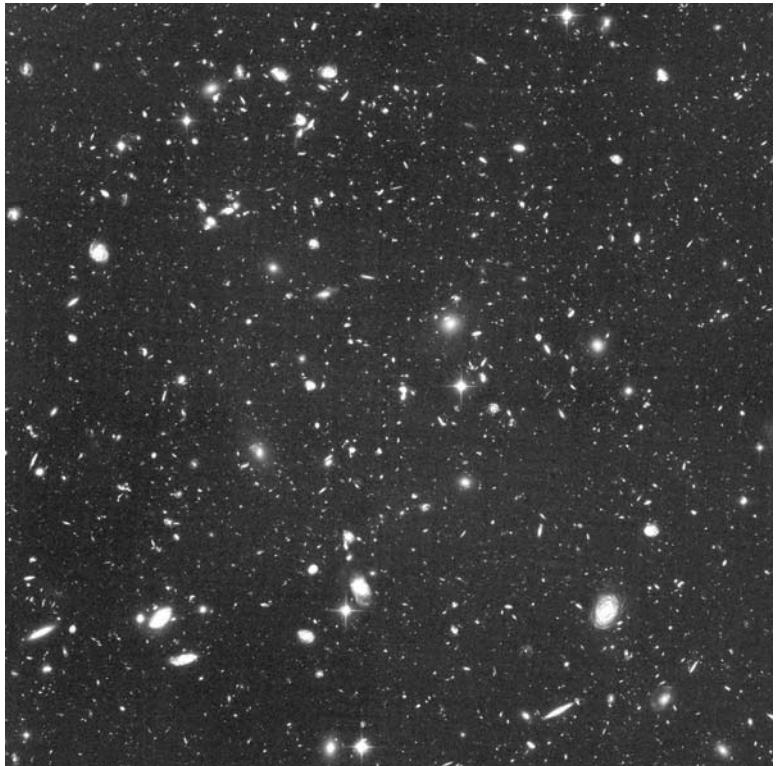


Fig. 15.2 The Hubble Deep Field image. This image was created by pointing the telescope into a relatively empty part of the sky and taking a very long exposure. Almost every smudge of light in the picture is a distant galaxy. (STScI/NASA.)

quasars are located in the centers of early galaxies. Many recent images of very distant quasars clearly show the surrounding galaxy, and some even display wisps of spiral arms surrounding the bright core. In other cases, especially among relatively nearby objects, quasars are members of clusters of galaxies, all with the same redshift. Nearly normal galaxies have been discovered at great distances, corresponding to very early times in the history of the universe. Thus galaxies as well as stars must have formed quite promptly after recombination.

The early appearance of stars and galaxies is a conundrum for modern cosmology. It might be expected that the formation of galaxies would require a significant passage of time after the primordial density enhancements first began to pull themselves together. The simplest models of galaxy formation have difficulty producing galaxies so early in the history of the universe, yet they exist. The *Hubble Space Telescope* is able to resolve galaxies at incredible distances, and some remarkable data have resulted. The most distant galaxies seem fragmentary and inchoate, as might be anticipated; anything we call a galaxy is, by definition, a fairly well-defined object containing highly organized structures such as stars, globular clusters, and possibly spiral arms. A little later than the era of these primitive objects, after less than 10% of the age of the universe had passed, galaxies seem to have formed, though not necessarily in their final configurations.

Galaxies are seen at ever larger redshifts

As we probe deeper and deeper into the past, we find evidence that the process of galaxy formation was anything but tranquil. Elliptical galaxies seem to have undergone collisions and mergers, triggering rapid star formation that used up much of the available gas while expelling most of the remaining gas and dust. However, elliptical galaxies also seem to have settled into their characteristic shapes rather quickly; for example, an object that is nearly certain to be an elliptical, and which is apparently in its final stages of formation, has been found at $z = 3.8$. Ellipticals apparently underwent a rapid evolution and then settled down into their present form. Some giant ellipticals, such as those found in the centers of great galaxy clusters, have continued to grow, apparently by swallowing up nearby galaxies in a process known as **galactic cannibalism**. Even in the recent universe, collisions and mergers are known to occur, especially in rich clusters that have strong gravitational attractions near their centers. Spiral galaxies too have led eventful lives, especially those spirals that inhabit denser clusters. Some spirals appear to have experienced intense and repeated bursts of star formation, perhaps triggered by near misses, or even outright collisions, with their neighbors. A few spirals were destroyed by such encounters, while others sprouted ringed and spoked arms marked by bright young stars. Even the apparently placid Milky Way has swallowed some of its dwarf galaxy neighbors.

The problem of galaxy formation and evolution is intimately tied to other issues in cosmology, including the initial conditions of the universe and its matter content. The prospects for genuine breakthroughs have never been better, however, and rapid progress is occurring on many fronts. Cosmology has historically been one of the most data-starved of sciences, an unfortunate circumstance that forced theory to advance almost blindly, with little help or discipline from observations. Recent advances in telescope technology, both ground based and space based, have drastically changed this situation; for the time being, observations have begun to gain the upper hand on theory. The most exciting times in any science always occur when qualitatively new data become available. In recent years cosmology has suddenly found itself in such a condition. Perhaps we are now on the verge of a truly thorough understanding of the contents of the universe.

The surprisingly violent history of galaxies

The matter content of the universe

Until relatively recently, cosmologists assumed that ordinary matter, that is, the elements that make up the Periodic Table, was the only significant constituent of the universe. This assumption was challenged by the gradual accumulation of contrary data. The first hint that something was amiss came from observations of the rotation speeds of spiral galaxies. The rotation speed remained constant or even increased at the edge of the galaxy even as the light dropped to zero. According to Kepler's law, the rotation speed would decrease with radius if most of a

Evidence for dark matter



Fig. 15.3 The gravitational field of the galaxy cluster Abell 2218 produces gravitational-lens images of an even more distant galaxy cluster. The images appear as arcs encircling the massive core of Abell 2218. (W. Couch, University of New South Wales; R. Ellis, Cambridge University; STScI/NASA.)

galaxy's mass was concentrated at its center. The observations showed that while the light decreased as a function of radius, the mass did not. In other words, there was a great deal of **dark matter** in the galaxies.

Clusters of galaxies provide abundant evidence for the presence of dark matter. Although clusters do not exhibit systematic rotation as does a spiral galaxy, the velocities of the member galaxies still depend on the amount and distribution of mass within the cluster. The observed velocities imply that substantially more mass is present than can be contained in the visible stars. Further support for the presence of considerable dark matter came from an early generation of X-ray satellites launched in the 1970s. Many galaxy clusters, it seems, encompass a significant quantity of X-ray emitting gas distributed throughout the cluster. The temperature of this gas is sufficiently high that it would escape if the cluster were not very massive and the gravitational potential correspondingly deep.

Particularly dramatic evidence of dark matter is provided by observations of gravitational lensing due to a galaxy cluster located between us and an even more distant galaxy or quasar. Examples include the striking lensing of distant galaxies by the galaxy cluster Abell 2218, as illustrated by Figure 15.3. The amount of lensing is determined to a large extent by the quantity of gravitating matter in the foreground galaxy cluster. Here again the implied mass is many times what can be accommodated by luminous matter. In terms of the density parameter Ω , the cluster data implied values in the range of $\Omega_M \approx 0.1\text{--}0.3$. Simply adding up the mass in visible stars, on the other hand, gives a value closer to $\Omega_{\text{stars}} \approx 0.005$.

Clearly, then, a considerable fraction, if not most, of the mass of galaxies and galaxy clusters is dark. While some of this dark matter could be in the form of baryons, for example interstellar gas and dust, there seems to be more mass present than can be accommodated by ordinary matter. Let us review the candidates for dark matter, including both ordinary baryonic matter and more exotic forms of dark matter.

Baryonic matter

The form of mass with which we are most intimately familiar is baryonic; that is, it is composed of ordinary atoms consisting of protons and neutrons. It has now become abundantly evident that the total density of baryonic matter is well below the critical density. Calculations of big bang nucleosynthesis, and the measured abundances of light elements, point to a value of $\Omega_b \approx 0.02h^{-2}$, which is well below unity and is even below the value inferred from galaxy cluster observations, for any reasonable estimate of the Hubble factor h . The amount of easily seen luminous matter, however, is much smaller than even this modest density. Much of the baryonic matter must therefore be dark. Is this baryonic dark matter in the form of gas, dust, compact dead stars, or something else? One place to seek the answer to this question is in the halos of spiral galaxies, which are known to contain considerable dark matter. Extremely long-exposure photographs of both spiral and elliptical galaxies have, in many cases, revealed faint light coming from greatly extended halos. This certainly confirms that at least some of the matter in massive halos is baryonic, since as far as we know only ordinary matter can emit light.

There are relatively few avenues for observing the contents of these halos directly, but there is a very interesting technique, based upon gravitational lensing, which is capable of detecting compact objects in the halo of our own galaxy. A gravitational lens effect can be produced by objects of any mass, no matter how puny; however, the smaller the mass, the more minuscule the bending of the light. Nevertheless, a compact object of even a modest mass can distort the light from a distant source in predictable ways. For instance, if a compact object were to pass in front of a background star, it would split the image into multiple images. These multiple images would be too small to resolve independently; the net result would be a temporary brightening of the background star (Figure 15.4). Such a gravitational effect is called *microlensing* in order to distinguish it from the lensing by a large object such as a galaxy cluster. The appellation simply refers to the type of lensing object; there is no qualitative difference in the mechanism. This class of hypothetical inhabitants of the galactic halo has been dubbed the **MACHO**, for MAssive Compact Halo Object.

Several collaborations have formed, named with quaint acronyms such as OGLE and EROS, to search for microlensing phenomena. One in particular, the MACHO project, provides a representative example of techniques and results. Although a compact object can hardly be expected to pass directly in front of any particular background star, constant monitoring of millions of background stars should find a few of these low-probability events. The Milky Way has two nearby satellite galaxies, the Large and Small Magellanic Clouds, that can conveniently provide such a background of stars. If a compact object in the halo of the Milky Way passed in front of any of the millions of stars in one of the Magellanic Clouds, it would produce a change in that star's bright-

Much of the baryonic mass in the universe is dark

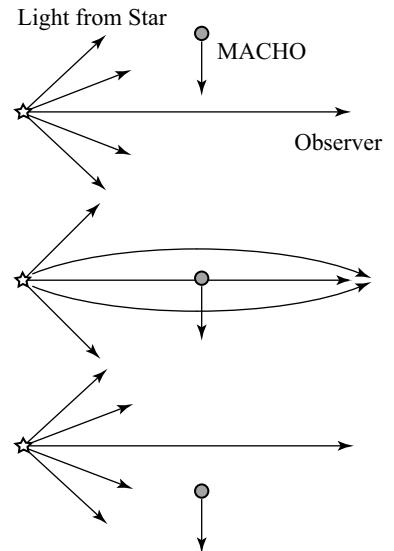


Fig. 15.4 Illustration of microlensing by a MACHO. As the MACHO passes in front of a more distant star, light will be focused by gravitational lensing, making the star appear to brighten. When the MACHO moves past, the background star reverts to its original appearance.

A hunt for MACHOs

ness in accordance with the properties of a gravitational lens. To search for such microlens events, night after night the MACHO project photographed the Large Magellanic Cloud, along with stars in the bulge at the center of the Galaxy, using a CCD detector array on a telescope in Australia. Computers scanned each night's photographs, looking for any changes in stellar brightnesses over time. After several years of search, a significant number of microlensing events were recorded. The results indicated that compact objects could account for perhaps 20–50% of the mass in the halo. No particular object is specified by the data, although the observations are consistent with objects of a mass of $0.5M_{\odot}$, which implies that the typical object is a white dwarf. The failure to find more such compact objects provides evidence that the dominant dark matter component in the Galactic halo is nonbaryonic.

Neutrinos

A small neutrino mass could have great cosmic significance

Many suggestions have been made for the identity of the possible nonbaryonic dark matter. One obvious candidate is the **neutrino**, which was thought to be massless until recently. At the time of the neutrino's discovery its mass was postulated to be zero because there was no particular reason known for it to have a small, yet still nonzero, mass. On the other hand, there is also no theoretical reason that the neutrino must have precisely zero mass. (This is in contrast to the photon, which must be massless for strict theoretical reasons, as experiment has verified.) As a dark matter candidate, the neutrino has the added virtue of being a particle that unquestionably exists. Neutrinos were produced in the big bang in numbers comparable to those of photons, roughly a billion neutrinos per baryon. This enormous population means that even if neutrinos have a rest mass no greater than a billionth of the rest mass of the proton, they would still make a significant contribution to the mass density of the universe. Specifically, neutrinos alone would close the universe if their mass is about $10^{-7}h^2$ times the mass of the proton. However, current experimental evidence places the neutrino's mass well below this critical value.

But cosmology is not the only, nor even the first, reason that physicists are attempting to measure the mass of the neutrino. Some of the strongest evidence for a massive neutrino comes not from any majestic theory of the structure of the universe, but right from our cosmic backyard. According to our current understanding of nuclear theory, the Sun and other stars should emit copious quantities of neutrinos. But because neutrinos interact so weakly with ordinary matter, they are very difficult to detect; only recently has neutrino astronomy, the direct observation of these neutrinos, become possible. The first solar neutrino detector was built in the late 1960s and now has taken several decades' worth of data. This detector consists of an enormous quantity of a chlorine-containing fluid commonly used in dry cleaning, buried deep underground in an

The development of neutrino astronomy

old gold mine, the Homestake, in South Dakota.³ Even though neutrino interactions are exceedingly rare, the sheer quantity of fluid in the detector means that occasionally a neutrino interacts with a chlorine atom, resulting in a signature that can be detected. The detector must be buried in order to screen out interactions with ordinary particles from cosmic rays. These particles, mostly high-energy protons, can produce events that would overwhelm and confuse the detector; however, cosmic rays cannot penetrate far into the Earth. Neutrinos, in contrast, could fly right through a sheet of lead several *lightyears* in thickness; a few hundred feet of rock is essentially transparent to them.

Subsequent neutrino detectors, several of which are now operating, follow the general principles of the Homestake experiment: look for rare neutrino interactions in a large quantity of some substance. Some detectors make use of the rare-earth element gallium. Another type of detector is filled with water, a substance that has the advantages of considerably less toxicity and expense than either dry-cleaning fluid or gallium. The water-based detectors use water of extreme purity; high-energy neutrinos striking electrons, and antineutrinos striking protons, lead to brief flashes of light that can be detected by photocells surrounding the fluid. Nearly all neutrino experiments are very difficult. Neutrino events are sufficiently scarce that instrumental effects can be significant. A major source of error is that the shielding by the Earth is never perfect, so some spurious interaction events still occur due to cosmic rays.

A significant motivation for the construction of neutrino detectors was to study neutrinos emitted by fusion processes in the core of the Sun. Ever since the Homestake experiment began to monitor the Sun over three decades ago, it has detected only about a third to half as many solar neutrinos as theory predicts. This result has been confirmed by more than one neutrino detector, so we can be confident that the effect is not instrumental. Since other facets of stellar theory are well supported by observations of the Sun, this discrepancy led theorists to postulate that the neutrino has a very small mass. This conclusion is based on the behavior of the different members of the neutrino family. The three known species of neutrino are called the *electron neutrino*, the *tau neutrino*, and the *muon neutrino*. The electron neutrino is by far the most abundant; moreover, most detectors are sensitive only to it. As it turns out, the solar data could be explained handily if some of the electron neutrinos were converted into tau or muon neutrinos on the journey from the Sun to the Earth, a phenomenon known as *neutrino oscillation*. More specifically, a flux of electron neutrinos emerging from the Sun will, after traveling through space for some distance, transform itself into a mixture of the different types of neutrinos. For technical reasons, neutrino oscillation can occur only if neutrinos have mass. Should

³Legend has it that the experiment's principal investigator, Ray Davis, received a complimentary truckload of wire hangers from a grateful dry-cleaning supply company, which assumed he was opening a huge cleaning establishment.

Neutrinos from a supernova

neutrino oscillation take place, then a detector sensitive only to electron neutrinos would be blind to the arriving tau and muon neutrinos and their passage would go unrecorded, even though they began their trip as potentially detectable electron neutrinos.

However, the Sun is not the only source of astronomical neutrinos. One of the most exciting neutrino observations was made when a supernova designated SN1987A exploded in the Large Magellanic Cloud, the nearby dwarf irregular galaxy. On February 23, 1987, a puzzling burst of neutrinos was seen over a 13-second interval by detectors in both Japan and the United States. This fusillade occurred approximately 20 hours before SN1987A was first spotted by a telescope operator in Chile. Neutrinos and antineutrinos should be emitted in copious quantities by a supernova, as the protons and electrons are squeezed into neutrons during the collapse of the core. Because of their weak interaction with matter, the majority of the neutrinos zip into space immediately after the core collapse, whereas the photons are emitted only after a shock wave reaches the surface of the star; thus neutrinos from SN1987A were seen first even though they travel at nearly the same speed as the photons. The detection of neutrinos from the supernova thus confirmed the basic theory of supernova explosions.⁴

More importantly for cosmology, the supernova gave an upper limit on the mass of the neutrinos. If neutrinos were massless, then they would travel at the speed of light. In that event, all the neutrinos from a supernova would arrive at the Earth at essentially the same time. If neutrinos have some rest mass, however, their speeds must be slightly less than c , and would be spread out over a range of values; not all the neutrinos would arrive at the same time. The actual spread of neutrino detection times for SN1987A suggested that neutrinos have a mass less than 17 billionths of the mass of the proton, less than the mass required to close the universe for Hubble constants greater than $50 \text{ km s}^{-1} \text{ Mpc}^{-1}$.

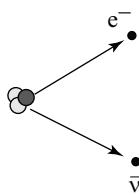


Fig. 15.5 When tritium decays it emits an electron and an antineutrino. Careful measurement of the energies of the emitted electrons provides limits on the mass of the neutrino.

Direct experimental evidence for neutrino mass has also been obtained from various laboratory experiments. Typically, these experiments attempt to infer the mass of the electron neutrino from observations of the radioactive decay of tritium, a heavy isotope of hydrogen. When tritium decays, it emits an electron and an antineutrino. A nonzero neutrino mass will reduce the upper limit on the portion of the decay energy that can be carried away by the electron. This is a very delicate and difficult measurement, but experiments so far suggest that the electron neutrino might have a mass of somewhere between 0.5 and 5 billionths of the mass of the proton. This is consistent with the supernova result, and again indicates that the neutrino alone cannot make up the bulk of the universe's dark matter. The experimental determination of the mass of all species of neutrino has been difficult, so this mass should be considered tentative but indicative.

⁴In 2002 Raymond Davis, the designer of the Homestake experiment, and Masatoshi Koshiba, the principal investigator for the Kamiokande neutrino detector, were awarded the Nobel Prize in physics for their pioneering work in observational neutrino astronomy.

Other recent experiments have placed new limits on the mass of the neutrino. The Super Kamiokande detector, a larger version of the experiment that detected the neutrinos from SN1987A, consists of a chamber located a kilometer underground, lined with sensitive light-detecting photocells and filled with 50 thousand tons of ultrapure water. Beginning in 1998, this detector obtained new evidence in favor of a nonzero mass for the neutrino. In this experiment, the neutrinos to be detected come not from a distant supernova, but from the Earth's upper atmosphere. The interactions of cosmic-ray protons with nuclei high in the atmosphere generate many types of particles that rain down toward the Earth's surface; among these particles are muon and electron neutrinos. Theory predicts that twice as many muon neutrinos as electron neutrinos should be produced by cosmic rays. The experiment counts the number of muon neutrinos and electron neutrinos (Super Kamiokande can distinguish these species) and compares the results to the expected (roughly) 2:1 ratio. Super Kamiokande found a shortage of muon neutrinos, which suggests that some of the muon neutrinos were converted into something else between the time they were formed and the time they passed through the detector. This provides direct evidence for neutrino oscillation, a phenomenon that had already been proposed to explain the shortage of electron neutrinos in the Homestake experiment. The data do not support the theory that the muon neutrino must be converted into the electron neutrino. The muon neutrino might be capable of transforming into the tau neutrino, which cannot be detected by Super Kamiokande, or it may become something else. Although the experiment provides evidence that muon and electron neutrinos have *different* masses, it does not determine with certainty what those masses are. The best evidence is consistent with a very small mass, with the heaviest neutrino having a mass no more than about one billionth of the mass of the proton.

Neutrino oscillation is evidence for neutrino mass

Although neutrinos may well prove insufficient to close the universe, their large cosmic number density means that with any nonzero rest mass they could still be an important component of the dark matter. Neutrinos are an example of **hot dark matter**. The epithet “hot” refers to the speed of the neutrinos as they move through the universe; massive neutrinos still travel at speeds close to that of light. The typical speed of a population of particles influences how they create structure in the universe as they begin to clump gravitationally. However, it has become clear that massive neutrinos cannot solve the dark matter problem. The observed distribution of galaxies cannot be explained by a picture of baryons dancing to the gravitational tune of massive neutrinos.

Cold dark matter

If it is true that $\Omega_M = 0.3$ but only about 10% of that mass is baryonic, then we must hypothesize the existence of exotic forms of matter. Any nonbaryonic matter in the universe, such as the exotic particles predicted by various theories of particle physics, must remain aloof from

*WIMPs and dark matter**Candidate dark matter particles*

all matter, including itself, except for its contribution to gravity; more direct interactions would have significant, and observable, effects upon the universe. Consequently, the type of particle hypothesized to account for nonbaryonic dark matter is called a **WIMP**, for *Weakly Interacting Massive Particle*.⁵

Many suggestions have been put forward for the identity of the possible nonbaryonic dark matter. One obvious candidate, already discussed, is the neutrino, but it seems that neutrinos alone do not possess enough mass. But neutrinos are by no means the only possibility for a WIMP. Other, more exotic, particles could contribute to the mass density of the universe. Such particles are required by various theories of particle physics; grand unified theories in particular provide no shortage of candidates. In contrast to the neutrinos, it is thought that these particles would have large rest masses, and would travel slowly through the universe. This would make them amenable to clumping together gravitationally into large regions in precisely the sort of way desired for dark matter. Because these hypothetical particles have low velocities, they are collectively called **cold dark matter**.

The spontaneous symmetry breaking that separates one fundamental force from another may require new massive particles called *gauge bosons*. The effect that may account for baryogenesis might be mediated by a very heavy particle called a *Higgs boson*, a hypothetical boson that could also play an important role in determining the masses of all particles. A type of GUT called supersymmetry postulates that every known boson or fermion must have a supersymmetric fermion or boson partner. The fermion partners are named by adding the suffix “-ino” to the name of the boson, while the boson partners are named by adding a prefix “s.” For example, the photon, a boson, has a fermion partner called a “photino”; the fermionic quark has a boson partner called a “squark.” The lightest of these supersymmetric particles, and the one that would be stable, is given the name *neutralino*, and it is a leading dark matter candidate. The Large Hadron Collider of the CERN laboratory, due to begin operation by 2008, may be able to detect this particle, if it exists.

It is possible, however, that some candidate WIMPs could be detected in the laboratory before they are seen in particle accelerator experiments. Like neutrinos, these WIMP candidates are expected to interact with ordinary matter only extremely rarely; nevertheless, they could scatter off atoms in a detector, releasing a photon that could be registered by a photocell. Since the Milky Way Galaxy is immersed in a great agglomeration of cold dark matter, such a detector could, in principle, see an annual variation in the detection rate and angle of incidence of such photons due to the Earth’s motion around the Sun and the Sun’s motion around the Galaxy. A number of experiments of this nature are currently operating or are proposed for the future. At the moment, nothing has been detected; the experiments provide only limits on the

⁵The previously mentioned dark matter candidate named MACHO was so dubbed partly in jest, to contrast with the WIMP.

types of WIMPs that might exist, and their interactions with ordinary matter. But physicists are hopeful that cosmology will provide valuable clues as they attempt to understand the fundamental nature of matter and energy. It may be that our best hope for examining theories such as GUTs will come from testing them with the evidence from the universe itself.

The origin of structure

If the universe is isotropic and homogeneous, how did the present large-scale structure form? The simplest answer is that structure is the result of gravitational attraction amplifying the small overdensities that were present at the recombination epoch and are seen imprinted in the CBR. These ripples in the early universe grew into the galaxies and galaxy clusters that we see today, but because the ripples were evenly distributed throughout the universe, on the large scale the universe is still homogeneous. Our study of the growth of structure must eventually address the origin of these fluctuations. A major area of current research is the study of what might have provided the seeds for the inhomogeneities observed in the CBR. Several ideas have emerged; we will consider one of these theories in more detail in the next chapter. For now, it is sufficient to know that the initial perturbations did exist: we see them directly in the *WMAP* results, and we infer them indirectly through the existence of structure in the universe today.

Perturbations can occur on many scales; perturbations on some scales will grow, while those on other scales will die out. How can we deal with all these different scales in a general way? Just as we can break light into its component wavelengths, so can we decompose any arbitrary density inhomogeneity into a spectrum of waves, each with a specific wavelength. We are interested in the behavior of these fundamental waves, or *modes*, of the perturbations. The spectrum tells us the amplitude of each mode. If we can determine the spectrum, and the evolution of each of its constituent modes, then we have all the information we need to describe the evolution of any arbitrary inhomogeneity. The *WMAP* results give us the amplitude of some of these modes at recombination; however, we cannot directly measure the spectrum of primordial seed fluctuations, so we must formulate a hypothesis. The simplest initial spectrum assumes that the ratio of the perturbation in density to the background density is a constant for every mode. The constant specifies the amplitude at the time when a particular mode's wavelength is the same as the Hubble length. This spectrum is not as *ad hoc* as it might seem; it is derived from a well-founded assumption about the random nature of perturbations. It also has the desirable feature that it is consistent with the anisotropies observed in the CBR. It is called the **Harrison–Zel'dovich spectrum**, after the astrophysicists who independently applied it to studies of structure formation. The Harrison–Zel'dovich spectrum requires that all perturbations have the same am-

The beginning of cosmic structure can be seen in the CBR

Decomposing perturbations into a spectrum of wavelengths

plitude when their size is the same as the Hubble length. The Hubble length, of course, increases with time. Hence the smaller the scale of any particular perturbation, the earlier the time at which it equals the Hubble length. A perturbation begins to evolve at the point in time when its wavelength first becomes smaller than the Hubble length.

During the radiation era, the strong coupling between radiation and matter prevented the perturbations from growing. Thermal equilibrium of baryons and photons ensured that the radiation interacted strongly with the matter. The photons were not yet freely streaming through the universe, but could only diffuse through the matter, in a process called *photon diffusion*. Photons in slightly hotter regions diffused into slightly cooler regions in order to maintain equilibrium; because of the tight coupling between the matter and radiation during the early universe, the photons tended to drag the matter with them as they moved. This phenomenon meant that the photons were able to squelch perturbations in the matter, an effect known as **photon damping**. The scale of this damping is essentially determined by how far a photon can travel, on the average, before scattering from matter. Perturbations with a scale smaller than this length are suppressed by photon diffusion. However, photon damping is of less importance if the universe is dominated by nonbaryonic dark matter, since such dark matter decouples from the radiation at a much earlier time than does ordinary matter.

Particles such as WIMPs or neutrinos are said to be *collisionless* because of their weak mutual interaction. Their behavior must be treated by statistical methods; such methods show, however, that even collisionless matter causes damping of perturbations, in a manner similar to photon damping. Collisionless, or weakly interacting, particles can diffuse from regions of high density to regions of low density, thus tending to smooth out perturbations. This is called **collisionless damping**, and as is true for photon damping, its scale is determined by the mean distance traveled by a particle over a given time interval. For free-streaming, weakly interacting particles, this length is set by the average speed of the particle. Hot dark matter particles, such as neutrinos, have velocities close to the speed of light and suppresses perturbations on much longer length scales than do cold, that is, slowly moving, dark matter WIMP candidates. Thus, hot or cold dark matter each leads to a characteristic type of structure formation. If small-scale perturbations can grow without being smoothed away by collisional or collisionless damping, then those small-scale perturbations can grow for a longer time than can larger-scale perturbations, which fall beneath the Hubble length at later times. Thus, if damping is not important, bottom-up structure formation should predominate, with smaller structures forming first. Damping, which would be present and strong for hot dark matter, leads to a top-down scenario, as the small-scale perturbations are quashed before they have a chance to grow.

Once the epoch of recombination is reached, the photons separate for good from the matter, imprinted with whatever spectrum of perturba-

The effect of damping smooths out perturbations on certain length-scales

The nature of the dark matter determines the size of the structures that form first

tions was then present.⁶ *WMAP* and other CBR anisotropy experiments can measure this perturbation spectrum out to a scale set by the resolution of the instrument. These scales are beginning to tie into the large scales of galaxy clustering. For example, the fundamental scale at the time of recombination was about 0.5 Mpc in size. Since recombination occurred near a redshift of $z = 1000$, this scale is now 1000 times larger, or 500 Mpc. Since the resolution of *WMAP* observations is about 0.2 degrees, its data can set important constraints on what is to be expected at even shorter scales, down to contemporary distances of 50–100 Mpc.

To connect the observed CBR fluctuations to the structure seen today, we must understand the evolution of matter after recombination. The gravity from the clumping dark matter draws baryonic matter into over-dense regions, and the self-gravity of the baryonic matter then speeds the process. The nonbaryonic dark matter is collisionless, so the size of the structures that it forms is set by its velocity, that is, whether it is hot or cold. Baryons, on the other hand, are collisional, and when they are pulled together into self-gravitating clumps they collide and shock, heat up, and then radiate away some of this energy, forming even smaller structures such as galaxies and stars. This happens relatively quickly; *WMAP* observations provide indications that the first stars formed very early, only a few hundred million years after the big bang.

The universe remains largely mysterious during the interval between recombination and the subsequent time, about one billion years later, that corresponds to redshifts at which we can begin to see galaxies. The CBR observations provide much valuable information, but the questions can also be approached from the other end, as it were, by examining the structure present today in the galaxies and galaxy clusters. We can then attempt to relate that data back to the structure present in the CBR at recombination. For such investigations, we need a thorough statistical sampling of galaxies that can only be obtained by detailed galaxy surveys.

Surveying the universe

Observations of galaxies and galaxy clusters provide a particularly valuable source of cosmological information. When we look into space we see only the luminous objects, such as the stars, galaxies, and quasars, not the dark matter that dominates the total mass of the universe. The luminous objects are our only probe of the gravitational evolution of the universe, both on the cosmic scale in the form of the Hubble flow, and on the local scale where effects such as the infall of mass into structures like galaxies, clusters, and superclusters must be included. How can we discern the shapes, sizes, and distributions of these cosmos-girding structures? Historically, the first approach used was simply to count the number and position of galaxies seen on photographic survey plates.

⁶Chapter 14 explains one mechanism by which this imprinting occurred.

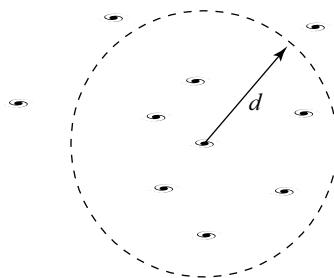


Fig. 15.6 The correlation function gives the probability of finding another galaxy within a distance d of a given galaxy.

The statistics of galaxy clustering

This was fairly tedious work, and the amount of information that could be compiled was restricted by both human and photographic limitations. But even this early work suggested that there is substantial galaxy clustering.

Since a firm quantitative measure of galaxy clustering is required, simply examining maps of galaxy positions is not sufficient. The human eye is easily fooled; the human brain is so attuned to pattern recognition that it has little difficulty in discerning patterns in what is really just pure random noise. It has often been remarked that this must reflect an evolutionary pressure; far better to see a leopard that isn't there, than to miss the one that is! Be that as it may, it renders our subjective judgements of galaxy clustering quite unreliable.

Quantification of the structures in the universe is not an easy matter. As always, there are many subtleties associated with data analysis of this kind, but the basic concepts are not too difficult to grasp. The goal is to measure the average probability that a randomly selected galaxy will lie within some distance d of another galaxy; such a probability is given mathematically by a **correlation function**. As remarked, the first attempts to measure galaxy correlation functions made use of photographs of the sky. Of course, photographs can show only the projection of the three-dimensional galaxy distribution upon the two-dimensional apparent surface of the sky. This may not seem like much with which to start, but it is possible to draw some inferences about the tendency of galaxies to cluster just from this information. Since the galaxies are photographed on the celestial sphere, without knowledge of the distance to the galaxy, their proximity on the photograph demarcates only their *angular* separation on the sky. In this case, we are interested in the probability of finding the image of another galaxy within a circle of some angular size. An angular size in the sky will correspond to some separation d on the physical photograph. Selecting any individual galaxy, we draw a circle of some specified radius d centered about that galaxy (Figure 15.6). We count the number of galaxies whose images are located on the photograph within the distance d from the chosen galaxy. When this procedure is carried out for all the galaxies in a given sample, we can compute the average number of neighbors of any galaxy within that fixed distance d . By repeating such a measurement for all d , out to some reasonable limit, we compute the average number of neighbors as a function of the distance of separation.

Clustering is not the only effect that causes two galaxy images to lie within a given distance on a photograph. Even if galaxies were randomly distributed on the sky, we would find some apparent clustering merely due to the inevitable occurrences of coincidental alignments. We can, however, take that into account mathematically; it is not all that difficult to compute the expected number of neighbors within any specified distance d for a random scattering of images across an area. Any clustering over and above what would be obtained from a random distribution is described by the *two-point correlation function*.

This sort of analysis of survey photographs of galaxies always yields a measured two-dimensional, two-point correlation function that deviates from a purely random galaxy distribution. This is the mathematical way of expressing what we have already stated in words: galaxies are more likely to be found near other galaxies; that is, galaxies cluster. For more information, we may progress to the *three-point correlation function*, which measures the probability that three galaxies will make a triangle of a given size. In principle, we could keep extending this process, but in practice there is little point in going beyond the four-point correlation.

The value of any of these statistical measurements depends upon the quality of the data. The number of galaxies is enormous, and gathering statistics on them is hard work. As an illustration of the dedication of astronomers, in the late 1960s C. D. Shane and C. A. Wirtanen counted a few hundred thousand galaxies, by hand, from photographic plates! Fortunately, such devotion to scientific duty is no longer necessary. The availability of automated telescopes, electronic CCD detectors, fiber optics, and high-speed computers has revolutionized the field of galaxy surveys and opened up a new era in the study of large-scale structures in the universe.

The most important improvement in galaxy surveys comes from obtaining information about the full spatial distribution of galaxies, rather than just their apparent distribution on the two-dimensional celestial sphere. From Hubble's law we know that redshift provides a measure of distance, and with better redshifts for more galaxies it becomes possible to construct extensive three-dimensional maps of the distribution of galaxies. (The third dimension is usually plotted as redshift, not as distance, since redshift is what is measured. This avoids model dependencies in converting redshifts to distance at large z .) Since the 1970s, the quality and quantity of redshift data have constantly improved. The quantitative analysis technique for galaxy redshift surveys is a simple extension of the two-point correlation function to three dimensions. Rather than measuring the two-dimensional *projected* distances between any two galaxies, we simply measure their three-dimensional separations, using the redshift as the third distance.

One of the first three-dimensional surveys was carried out by the Harvard Smithsonian Observatory in the 1980s. The redshifts to thousands of galaxies were painstakingly measured to form a three-dimensional view of a section of the local universe. This first view revealed striking features: galaxies apparently arranged in superclusters that appeared as long sheets surrounding vast empty regions known as **voids**. The success of this survey led to the development of techniques to acquire ever greater numbers of redshifts. In the 1990s the redshift survey of the Las Campanas Observatory in Chile measured redshifts for tens of thousands of galaxies to distances of 100–200 Mpc (Figure 15.7). The 2dF Galaxy Redshift Survey has measured over 100,000 galaxy redshifts, and the Sloan Digital Sky Survey is underway in pursuit of the ambitious goal of measuring redshifts to one million galaxies, while determining the

Early clustering studies

Modern sky surveys

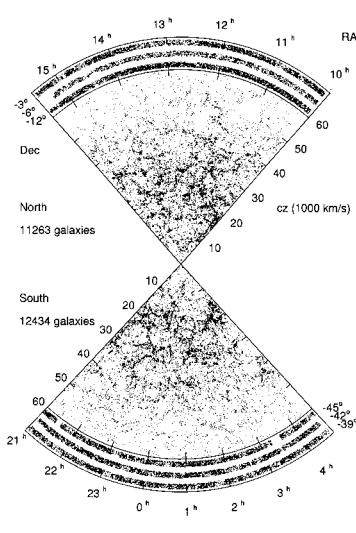


Fig. 15.7 Two slices through the universe, from the Las Campanas Redshift Survey. The distance is given by redshift, which in this survey goes as high as $z = 0.2$. Galaxy clusters tend to appear as long strings of galaxies directed radially toward us because a significant component of their redshifts is due to peculiar motions of the galaxies. This causes a distortion in the appearance of the cluster. The number of galaxies in the survey decreases at large redshift, as galaxies become fainter and more difficult to detect. (Lin et al. 1996.)

Distortions due to peculiar velocities

position and brightness of over 100 million celestial objects, including stars and quasars as well as galaxies.

The capabilities of these galaxy-redshift surveys have improved remarkably in the last two decades. What makes such projects possible are modern electronic CCD detectors, which form a digital image that can be stored in and analyzed by computers. Redshifts require spectroscopic measurements for each galaxy, and that involves more effort than taking a simple photograph of a region of the sky. In the past, it was necessary to obtain redshifts one by one, but now they can be acquired for a large number of galaxies at once. The Sloan Digital Sky Survey accomplishes this by constructing a special aluminum plate that fits into the focal plane of the telescope. Within a chosen field of view will be a large number of galaxies. A hole is drilled into the plate at the precise location of each galaxy on the image plane. Next, fiber-optic cables are connected to the holes, with each cable leading to a separate spectrograph. In this way the Sloan Survey can obtain up to 640 simultaneous galaxy spectra. On a good night's observing, a number of such plates can be used to measure thousands of redshifts. Thus it becomes feasible to collect a million galaxy redshifts with only a few years of very hard work!

From data giving a three-dimensional view of galaxy distributions, it is possible to compute the extent of clustering at different scales. This yields a power spectrum similar to that found by *WMAP* for the CBR fluctuations, but this time the fluctuations are represented by the galaxies themselves. To determine the degree of clustering at some scale, such as 10 Mpc, the computer constructs a sphere 10 Mpc in radius and counts the number of galaxies contained within that sphere. This operation is performed for every point in the data set; the larger the variation in number of galaxies contained within that volume, the greater the amount of clustering on the 10 Mpc scale. This procedure is repeated for each scale to obtain the full power spectrum.

Galaxy surveys and cluster statistics can provide valuable information about the evolution of the universe and the structure therein, but there are a number of complications in measuring the true three-dimensional structure of the universe. One such complication arises because we use redshift as a measure of distance. But we cannot distinguish redshifts due to the expansion of space, that is, the Hubble recession, from the peculiar motions of individual galaxies. An example of an effect due to peculiar motions is illustrated in Figure 15.7. All the members of a specific galaxy cluster should have approximately the same cosmological redshift and should appear as a compact cluster in the plot. The individual galaxies also have peculiar velocities, however, and these add to or subtract from the redshift measured. The result is that the appearance of the cluster on the plot becomes elongated in the z direction. This phenomenon has been dubbed the “Finger of God,” because the

cluster members lie upon a line pointing directly at the Earth.⁷ Another effect arises if the peculiar motions of the galaxies are systematic rather than random. If galaxies are falling toward the center of a cluster, then the galaxies on the far side of the cluster will have lower redshifts and the galaxies on the near side will have higher redshifts, resulting in a compression of the cluster in redshift space. Understanding these distortions is important because they result from clustering itself, the very phenomenon we seek to study. One method of accounting for the artifacts due to peculiar motions makes use of the fact that peculiar-velocity effects shift a galaxy's position in the radial, but not the angular, direction. The correlation function computed exclusively in the angular directions on the sky can be compared with the correlation function computed in the radial (redshift) direction. Systematic differences between the two correlations allow the amplitude of the distortions due to peculiar motions to be estimated.

The samples themselves also contain inherent limitations. Every galaxy survey is inevitably biased in some way. A perfect survey would accurately measure the redshift for every galaxy, no matter how luminous, in a specified large volume of the universe. Such an ideal survey would be said to be *complete*. Needless to say, this is difficult to achieve. The farther we go in redshift, the more we bias our sample toward the bright galaxies, those we can see for large distances. On the other hand, a galaxy survey that is complete, or as complete as humanly and technologically possible, for all galaxies down to some particular minimum brightness cannot probe to a very large redshift, although as telescopes and detectors improve that limitation is diminishing. Moreover, the appearance of galaxies as seen through some specific filter will change as z increases, as the filter samples different rest wavelengths of the galaxies' light due to the redshift. Galaxy evolution can also distort the results, especially as we go to greater and greater distances; the luminosity of very young galaxies may not be well described by the appearance of nearby, and hence much older, galaxies.

When we carry out surveys of galaxies to study structure in the universe, we are faced with another significant uncertainty: do the galaxies cluster in the same way as the dark matter? In other words, is the distribution of the luminous matter throughout space really the same as that for the unseen dark matter, or is there **biased galaxy formation**? If the galaxies we observe are not representative of the overall matter distribution, even in baryons, then the statistical properties we derive from studying them may have little relevance to the large-scale structure of matter. The motivation for studying biased galaxy formation arose originally from the observation that the mass-to-light ratio in galaxy clusters implied $\Omega \approx 0.2$, much less than the critical value of unity. If $\Omega_M = 1$, then most of the mass in the universe could not be associated with galaxy clusters and, by extension, luminous matter. An example

A question of bias

⁷This would surely violate the Copernican principle if it represented a real structure.

of how such a situation could occur is provided by *high peak bias*. In this picture, the dark matter is spread out fairly evenly through space, with a wide range of density fluctuations superposed on the overall distribution. Ordinary baryonic matter collects preferentially in regions of maximum dark matter density, and galaxies form only in the highest peaks in the dark matter density perturbations. The resulting galaxy clusters are then not representative of the overall matter distribution.

With the development of the concordance model of the universe, it now appears that the universe is flat, $\Omega = 1$, but only a fraction of that total is in the form of mass, namely $\Omega_M = 0.3$. The inferred mass density from mass-to-light ratios in clusters is therefore roughly consistent with the overall mass density of the universe. This suggests that substantial biased galaxy formation is not required. At least at scales for which we have the best data, approximately up to $40h^{-1}$ Mpc, the observations are well fitted by a cold dark matter model with values of the cosmological parameters consistent with the concordance model.

Apparently the most luminous galaxies and the clusters with which they are associated do trace the overall mass distribution of the universe. This is not to say, however, that all the different *types* of galaxies are distributed in the same way throughout the universe. To take one example, what if the formation process for bright galaxies is different from that for fainter dwarf galaxies? In that case, the clustering properties of bright galaxies might be entirely different from those of dwarf galaxies. Substantial evidence does exist that the distribution statistics of different types of galaxies are biased relative to one another. Elliptical galaxies and highly luminous galaxies are found more frequently in the centers of rich clusters than are spiral galaxies. The most common type of galaxy overall appears to be dwarf galaxies. Within our Local Group are several dwarf galaxies, some of which contain scarcely more than a million or so stars. Seeing faint dwarf galaxies at any significant distance is difficult, but there is evidence that faint dwarf galaxies are less clustered than bright galaxies.

Clearly, considerable uncertainty remains in our understanding of the process of galaxy formation. We know that the density perturbations were not uniform at the time of structure formation. Some regions would have been *very* overdense, compared to the mean, while others might have been quite underdense. The most overdense (and underdense) regions would have been the rarest, and bright galaxies might have preferentially formed in the anomalously dense regions. Less overdense regions might have collapsed after these peaks of density, while those regions only a little denser than the mean, which might be expected to be the most common, could have collapsed very late to form a relatively smooth background of extremely dim, or perhaps even dark, galaxies that we cannot detect at the present time.

Given the difficulties associated with any sample, it is important to perform as many different types of survey as possible. For example, additional information can be derived from surveys in frequency bands other than the visible. The largest catalog of galaxies in the far infrared



Fig. 15.8 An image of 1.6 million galaxies in the local infrared universe, obtained from the Two Micron All Sky Survey (2MASS). (Image from 2MASS, a joint project of the University of Massachusetts and the Infrared Processing and Analysis Center/California Institute of Technology, funded by NASA and NSF.)

band is provided by the *IRAS* satellite data. This survey favors objects that are bright in the infrared; a galaxy bright in the visible is often not bright in the far infrared, and, typically, *vice versa*. It also happens that the galaxies that shine brightest in the far infrared are often engaging in rather violent interactions of one form or another, such as collisions with nearby galaxies, leading to rapid and strong star formation. Such galaxies show a few bright, strong features in their spectra, making it easy to determine their redshifts.

The Two-Micron All Sky Survey (2MASS) is the first complete all-sky survey carried out in the near-infrared band. In comparison to the *IRAS* data, the 2MASS survey has better resolution and greater sensitivity. The 2MASS data are complete in bright galaxies out to $200h^{-1}$ Mpc. Because near-infrared wavelengths are less sensitive to newly formed stars than are the wavelengths used in the *IRAS* survey, 2MASS is not biased in favor of galaxies with many young stars. Near-infrared luminosity is proportional to galaxy mass; thus the near-infrared data provide a mass-selected survey of galaxies. Another advantage of the 2MASS survey is that its wavelength band is less affected by dust absorption; thus it covers the entire sky, including the region seen through the Milky Way. Optical surveys such as the Sloan survey are necessarily confined to only a portion of the sky, because dust absorption blocks the view near the plane of the Milky Way. Because 2MASS does not probe as far into space as do optical surveys, it can sample only the local universe, but statistical analysis of the data indicates that the 2MASS galaxies cluster in the same manner as galaxies observed in other wavebands.

Similar maps have been made for radio galaxies, with similar caveats. Radio galaxies, and other radio sources, can often be seen for enormous distances, making them good indicators of the distribution of galaxies at the largest scales; many radio sources are approximately a Hubble length away, which for practical purposes is as far as we can see. The maps of radio sources do not present any obvious structures to the eye; the data from these surveys are consistent with overall homogeneity on the largest scales.

The higher-energy portion of the electromagnetic spectrum can also be studied. Orbiting satellites have opened the X-ray universe to observation, and X-ray selected cluster surveys have been carried out. The cores of massive galaxy clusters are filled with hot X-ray emitting gas and so appear bright to an X-ray telescope. Because this hot gas would escape from the cluster unless it was held in by a strong gravitational field, the inferred temperature and density of the gas provides a measure of the cluster mass. Thus, X-ray surveys select for the most massive clusters, and provide another way to measure the mass density of the universe. And because such clusters tend to be bright in the X-ray band, X-ray surveys can, in principle, sample the universe to redshifts greater than $z = 1$ and beyond.

Simulating the universe

Structures in the universe emerged from an initial spectrum of seed perturbations formed early in the big bang. Their subsequent behavior was determined by gravity, and by other factors such as the type and density of dark matter, the baryon density, the expansion rate, and so forth. How can we test the complicated physics involved in structure formation in order to tie together the theoretical ideas with the observational reality? One approach is to construct a model of the universe that incorporates the details of the chosen theory, evolve it forward in time, and compare the results of such a *simulation* to observations such as the redshift surveys of galaxies. We can then ask whether the outcome of a given simulation resembles the observed present-day universe. If not, then some of the basic assumptions must be incorrect.

Detailed cosmological simulations do not consider the full Hubble sphere, but work instead with a comoving volume of the universe that is large enough so as to be representative; that is, it is a homogeneous volume. The conclusions derived can then be extended to the entire observable universe with the help of the cosmological principle. Typical sizes for simulation boxes correspond to around 100 Mpc per side in the present universe. Even with this restricted volume, cosmological models are sufficiently complex that their equations must be solved numerically with the aid of the fastest available computers. The recent dramatic increase in computing power has made a significant impact on many fields of astrophysics, particularly cosmology; greater computing capabilities mean greater spatial resolution in the simulation, and the inclusion of more complex interactions. These improvements lead to more realistic simulations whose results can be interpreted with greater confidence.

Many cosmological simulations use an *N-body* algorithm, in which discrete particles are placed into a volume and allowed to interact according to specific rules that model the appropriate physics. The simplest sort of system of this type simply regards the particles as point masses interacting gravitationally. In cosmological simulations, discrete massive particles represent the constituents of the model universe: ordinary

Putting theory to the test

The universe in a box

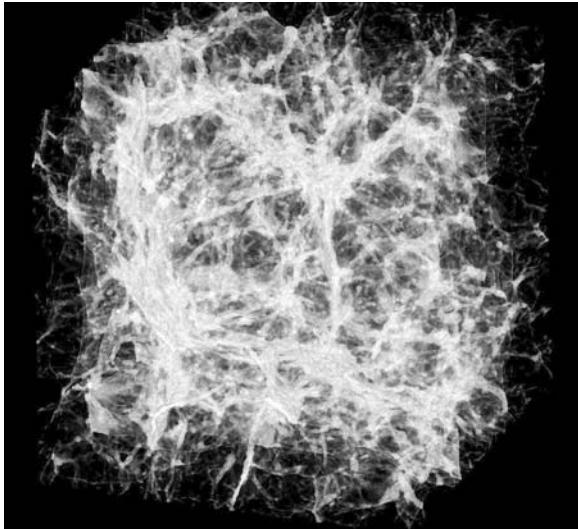


Fig. 15.9 A cosmological simulation showing density in a model that is consistent with the concordance model. (M. Norman, CASS/UCSD & P. Paschos, UCSD.)

matter, cold dark matter, and hot dark matter as appropriate for the model under study. The more particles, the better the resolution of the simulation. Some of the more recent simulations also include the effects of collisional gas dynamics in the ordinary matter, even though such interactions add considerably to the length and complexity of the computation.

A typical simulation begins at an early time in the universe, when the density perturbations in the matter are still relatively small. The evolution begins and the matter particles start to interact gravitationally. The model is evolved forward in time until the present, where it can be compared with the observations. The large-scale structures that form in the computer model can be treated as if they were data from astronomical surveys, and various statistics such as two-point correlation functions can be computed. These analyses determine whether the assumed perturbations, matter content, and other conditions can reproduce the observed structure in the physical universe. We have only one universe to observe, but cosmological simulations provide a means to examine the consequences of different initial conditions in model universes.

Cosmological simulation models are classified according to the type of matter they include. For example, the hot dark matter model assumes that the dominant form of nonbaryonic matter is a neutrino, or something similar, that has an extremely high velocity early in the universe. The characteristic high velocities of hot dark matter lead to very large structures, because the collisionless damping length is long. In neutrino dominated hot dark matter simulations, the first structures to form are superclusters. The neutrinos create the initial structures, and baryons fall toward the gravitational attraction of the neutrinos. Once the baryons begin to collapse, they flatten rapidly into pancakes, which then fragment into smaller objects due to friction and heating of the baryons. The distribution of neutrinos, however, retains the large-scale

A hot dark matter universe

structures. In the hot dark matter scenario, the massive neutrinos remain much less clustered than the baryons.

Although the hot dark matter model can reproduce some characteristics of the largest-scale structures, including filaments and voids, it has a number of features that render it unacceptable as a description of the actual universe. In order to fit the observed clustering properties of galaxies, the collapse must have occurred much too recently, at cosmological redshifts of approximately $z = 1$. This contradicts the observations of normal galaxies at redshifts much greater than unity, as large as $z = 10$, as well as leaving little accommodation for quasars, which are also seen at high redshifts. Hot dark matter models also tend to show excessively strong clustering; maps of galaxies created by computer simulations of hot dark matter show tight blobs of matter and huge voids. They simply do not match the statistics gleaned from current galaxy surveys.

A cold dark matter universe

A more successful model uses cold dark matter (CDM). Recall that cold dark matter consists of a weakly interacting particle that is very massive and is already quite cool (that is, has a low velocity) when it decouples from the rest of the contents of the universe. Because these WIMPs have such low velocities, the collisionless damping length is extremely short, which allows small-scale perturbations to grow unimpeded. Thus, in the cold dark matter model the first structures to form are smaller than galaxies. These objects then interact through gravity to form larger structures, pulling one another together and even merging. Eventually, galaxies and galaxy clusters result. This is more consistent with data indicating that galaxies formed quite promptly after the beginning of the matter dominated era, and only later were drawn into clusters and then superclusters. Our own Local Group may be an example; the Milky Way is certainly very old, probably about as old as any galaxy. Yet the size of the Local Group and the peculiar motions of the galaxies within it suggest that it is still forming and collapsing. Similarly, the Local Supercluster appears to be a fairly recent aggregation.

The CDM model has many attractive features. It is quite successful at reproducing the observed clustering properties, such as density and distributions, of galaxies. It accounts for the formation of galactic halos as a natural outcome of the collapse of the weakly interacting particles. However, simulations found that if the density of cold dark matter was great enough to account for a flat geometry, that is, $\Omega_M = 1$, the agreement with observations became less compelling. The CDM scenarios that best fit the observations have Ω_M closer to 0.3, consistent with the concordance model. This version, currently regarded as the most promising, is the Λ -CDM model.

Both hot dark matter and cold dark matter models must necessarily be highly simplified, due to the limitations of computing capabilities. For example, each computational particle represents an enormous quantity of mass even when millions of such particles are included. Moreover, in many simulations only the weakly interacting particles are evolved, but what we actually observe are the baryons and the galaxies they form. Adding baryons to the dark matter magnifies the difficulties and

boosts the demand on computational resources. Baryons possess much more complicated interactions than do WIMPs; they collide, heat, behave as fluids and develop shock waves, and so forth. The equations of hydrodynamics and thermodynamics have long been solved numerically on computers, but that is itself computationally challenging, requiring considerable spatial resolution for realistic solutions. It is very difficult to follow all the length-scales in a combined baryon-dark matter simulation, from the full 100 Mpc box down to the kiloparsec scales of galaxies and the even smaller scales associated with star formation. Recent efforts have dealt with this problem by making use of subdivisions of the cosmological grid; these subgrids follow the collapse of baryons onto ever smaller scales. Such models are providing new and detailed insights into the process of galaxy formation.

The complement and arbiter to the theoretical studies is improved observations. Satellite observations have already placed strict limitations on the size and spectrum of primordial temperature fluctuations. Future observations will push this to smaller and smaller scales, providing greater information about the state of the universe at the time when galaxies were just beginning to form. Ever larger galaxy surveys are providing detailed information about structure in the universe today. Exciting new observations will continue to be made. The *Hubble Space Telescope* continues to obtain deep images of the universe at high redshift. Its photographs have reached primordial gas clouds from which young galaxies may be just emerging. The planned successor to the *HST*, the *James Webb Space Telescope*, holds great promise for cosmology. This proposed telescope would consist of an 8-meter mirror optimized for the infrared. It will carry out deep imaging and surveys of galaxies at high redshift, when galaxies are first forming and evolving. We are on the verge of seeing the complete history of the assembly of galaxies laid out before us. Galaxies speak to us of the earliest times of the universe, but we cannot yet fully understand what they are saying. Only further research can clarify the mysteries of the galaxies.

Prospects for the future

Chapter Summary

Galaxies are the main visible constituent of the universe. Most galaxies are members of gravitationally bound clusters, which range in size from groups of a few galaxies to enormous rich clusters containing thousands of galaxies. The Milky Way is a member of a small cluster with a few dozen members, dominated by itself and the Andromeda Galaxy; this cluster is called the Local Group. The nearest large cluster to the Local Group is the Virgo Cluster, an irregular cluster with a few thousand members; as is typical for many large clusters, a giant elliptical galaxy

lies at its center. Structure does not stop at the level of galaxy clusters, but extends to superclusters, huge associations of matter extending over millions of parsecs. The Local Group and many similar small clusters are satellites of the Virgo Cluster and are falling toward it; together these clusters make up the Local Supercluster. Even larger aggregations of matter could be influencing the Local Supercluster gravitationally.

Structure forms because of the gravitational attraction of the matter content of the universe. Most of this mat-

ter is nonluminous dark matter. The nature of this unseen dark matter is one of the most important outstanding cosmological problems. Galaxies, including the Milky Way, appear to be surrounded by huge spheroidal dark halos of unknown composition. Observations have indicated that some of the halo mass must take the form of compact objects called MACHOs. At larger scales in the universe, clusters are found to represent great accumulations of dark matter. Estimates of Ω_M inferred from these clusters range from 0.1 to 0.3, far in excess of what can be accommodated by the abundances of primordial elements such as helium and deuterium. Thus we conclude that some 90% of the matter of the universe is not only invisible, but is nonbaryonic. It must consist of some type of weakly interacting massive particle, since anything that interacted more strongly would have significant observational and physical consequences. One obvious candidate for such a WIMP is the neutrino. If neutrinos have even a small mass, they could add considerably to the cosmic matter density because they are approximately as abundant as CBR photons. Data from neutrino detectors now show that the neutrino actually has an extremely tiny mass, but the experimental limits on the mass preclude it from providing enough matter density to be the main component of the dark matter. More exotic particles, of a type called cold dark matter because of their generally slow velocities, have been proposed. As yet, however, none has been detected, so the nature of the major matter constituent of the universe remains mysterious.

The origins of structure lie within the earliest moments of the big bang. The properties of the dark matter determine how the structure forms. Gravity acts to increase the amplitude of matter perturbations, but various damp-

ing effects tend to smooth the perturbations away. The results of these complex interactions are imprinted on the CBR as small temperature fluctuations. Space-based observatories such as the *WMAP* satellite have recorded these anisotropies in the CBR, making it possible to study events from the recombination epoch.

The properties of cosmic structure are measured by galaxy surveys. Modern large-scale surveys have obtained positions and redshifts of hundreds of thousands of galaxies. The structure observed in these surveys provides the standard against which cosmological models may be tested. Models of structure formation are complex and must be solved computationally; the results can then be compared to observations. Hot dark matter models assume some high-energy particle, such as a massive neutrino. However, the structures created by models of this type are too large and form too late in the history of the universe to be consistent with observations. At the opposite extreme are cold dark matter models, which assume a heavy, slow particle is dominant. Cold dark matter models form small structures, such as galaxies, first; these smaller structures develop quite early in the history of the model. This seems consistent with recent observations of apparently normal galaxies at high redshifts. Indeed, evidence from the *WMAP* satellite indicates that the first stars may have formed only 200 million years after the big bang.

Structure formation and evolution in the universe is one of the most active fields of current cosmological research. Larger surveys, new telescopes and satellites, and greater computer power all provide the hope that answers to many cosmological questions will be forthcoming in the foreseeable future.

Key Term Definitions

galaxy cluster A group of galaxies that are mutually gravitationally bound.

supercluster A cluster of galaxy clusters.

top-down structure formation The formation of large structures, such as galaxy superclusters or perhaps even the vast filaments and voids, prior to the formation of smaller structures such as individual galaxies.

bottom-up structure formation The formation of small structures first in the universe, perhaps galax-

ies or even smaller substructures, followed later by larger structures.

dark age The era, lasting hundreds of millions of years, between the epoch of recombination and the onset of star and galaxy formation.

reionization The point in time early in the universe, but after recombination, when the first stars formed and their ultraviolet light began to ionize the neutral hydrogen gas that filled the universe.

galactic cannibalism The process of galaxy merger in which a large galaxy disrupts and assimilates a smaller galaxy.

dark matter Any gravitating mass that does not produce light. Also often used in a more restrictive sense to refer to nonbaryonic gravitating mass, as in hot dark matter or cold dark matter.

MACHO Massive Compact Halo Object. Any object such as a white dwarf, neutron star, or black hole that could account for some of the dark matter in the halos of galaxies.

neutrino Any of three species of very weakly interacting lepton with an extremely small mass.

hot dark matter A type of nonbaryonic particle that has a velocity close to the speed of light at the time it decouples from other matter in the early universe. A predominance of hot dark matter would produce mainly large-scale structure, leading to top-down structure formation.

WIMP Weakly Interacting Massive Particle. A particle with a nonzero mass that participates only in the weak interaction.

cold dark matter A type of nonbaryonic particle whose energy is low at the time it decouples from other matter, and whose mass plays a key role in cosmic structure formation. Cold dark matter leads to bottom-up structure formation.

Harrison–Zel'dovich spectrum A proposed spectrum for the matter perturbations in the early

universe which later became the observed structure. The Harrison–Zel'dovich spectrum is scale-free; that is, perturbations of all sizes behave in the same way.

photon damping The tendency of photons in the early universe to smooth out inhomogeneities in matter with which they are in thermal equilibrium.

collisionless damping The tendency of weakly interacting (collisionless) matter to smooth out gravitational perturbations by freely streaming from over-dense to underdense regions.

correlation function A mathematical expression of the probability that two quantities are related. In cosmology, the correlation function indicates the probability that galaxies, or clusters of galaxies, will be found within a particular distance of one another, thus providing a quantitative measure of the clustering of galaxies (or of clusters).

void A huge region of space that is unusually empty of galaxies. Voids are not entirely empty, but they are underdense and contain far fewer bright galaxies than average.

biased galaxy formation The theory that the distribution of galaxies is not representative of the overall matter distribution in the universe because galaxies form preferentially from anomalously over-dense dark matter perturbations.

Review Questions

- (15.1) Explain the difference between hot dark matter and cold dark matter. Which one corresponds to top-down and which to bottom-up structure formation, and why? Describe the strengths and weaknesses of each corresponding model of structure formation. What other possibilities might exist?
- (15.2) What evidence do we have that there is substantial dark matter in the universe? Why is it not possible that such matter could be baryonic?
- (15.3) Assume there are one billion neutrinos for every proton. Using the limits on baryon density given by big bang nucleosynthesis, and assuming a Hubble constant of $70 \text{ km s}^{-1} \text{ Mpc}^{-1}$, how massive would the neutrino have to be, compared to the proton, to provide the balance of the closure density, $\Omega_M = 1$?
- (15.4) To what angular size on the CBR does a present length of one Mpc correspond?
- (15.5) What effects do peculiar velocities have on galaxy distributions in redshift surveys?
- (15.6) What is meant by biased galaxy formation? How can this effect complicate the interpretation of galaxy redshift surveys?
- (15.7) Suppose you are conducting a galaxy survey. What systematic effects will influence your sample as you go to higher redshifts?

This page intentionally left blank

The Inflationary Universe

16

Key Terms:

[I]t may happen that small differences in the initial conditions produce very great ones in the final phenomena.

—Henri Poincaré

The big bang model of cosmology is spectacularly successful as a scientific theory. Until quite recently, it would have seemed unthinkable that any physical model, much less one so elegant and simple, would enable us to begin to understand the universe as a whole, with all its daunting complexities. The big bang model makes definite predictions about observable quantities such as elemental abundances, and those predictions have been borne out to a remarkable degree. The agreement with observations is so good that the big bang model is the benchmark by which more sophisticated theories are evaluated; in many cases, cosmology has put rather severe constraints upon particle physics, guiding the development of our understanding of conditions we may never be able to reproduce on Earth.

Nevertheless, many questions remain to be answered. We still have little information about the properties of the main denizens of the universe. Mass may not be the dominant dynamical constituent of the universe, and ordinary baryonic matter is not even the major component of mass. If the universe is accelerating, as recent observations indicate, we have almost no knowledge about the mysterious dark energy that produces that acceleration. At an even more fundamental level, we have assumed all along that the universe is, on its largest scales, smooth, isotropic, and homogeneous, and have found good evidence that this is true. But *why* should it be so? There is no apparent physical compulsion for isotropy or homogeneity. Moreover, at smaller scales, the universe most certainly is neither isotropic nor homogeneous. Galaxies are not identical; beyond that, they form clusters, and even the clusters are organized into super-clusters. The better our telescopes become, the farther we can see, and the larger the structures that seem to appear. At what point does the assumed isotropy of the universe, so well measured by the *COBE* observations, begin? How can we explain the origin of structure within an isotropic universe?

Big bang models based on the Friedmann equation (11.22) are fully self-consistent, and difficulties such as these do not represent a failure of the models. Rather, the difficulties stem from the assumptions that were built in from the beginning, and from the incompleteness of our

Limitations of cosmological models

knowledge. For example, our models *assume* isotropy and homogeneity, and thus are not able to explain how isotropy and homogeneity could develop. Our incomplete understanding of particle physics, and our inability to perform experiments beyond an energy scale corresponding roughly to that of the hadron epoch of the early universe, further limits our knowledge. We could abandon the effort, and claim that this is the best theory of the universe that we can construct. This is an unsatisfying resolution, however; we have been able to advance as far as we have come in cosmology by insisting that the universe is knowable. Therefore, let us ask the difficult questions, and begin to seek explanations for them. Even if our first efforts may prove to be faulty, perhaps they will point the way to better explanations.

Unresolved issues in the big bang model

The horizon problem

One of the most nagging of the unsolved mysteries has to do with the very homogeneity and isotropy that has been our guiding principle in the creation of cosmological models. The homogeneous and isotropic Robertson–Walker metric is an exceedingly special solution to the Einstein equations, valid for conditions of high symmetry. We should perhaps be surprised that the real universe seems to be consistent with it. After all, the most general solutions to the Einstein equations, even absent the cosmological constant, are neither isotropic nor homogeneous. There is a fairly large class of known solutions, called the *Bianchi models*, that are homogeneous but not isotropic. The Robertson–Walker metric, however, is unique; it can be proven that this metric is the only solution to the Einstein equations for an isotropic, homogeneous universe. Why, out of the large number of more general solutions available, did the universe choose one so special? The measurements of the CBR by *COBE* showed that the universe must be isotropic to better than one part in 10^5 . If this is such a special condition, why does it just so happen that the one and only universe is isotropic?

Another mystery is how the universe even knows that it is homogeneous and isotropic. The present observable universe is so large that light requires billions of years to cross it; how can two points separated by so much distance resemble one another at all? We might be inclined to postulate that since the observable universe was initially small, it was once easy for particles to interact and thus for an overall equilibrium to prevail. A homogeneous portion of universe then simply enlarged due to the universal expansion, explaining the observed smoothness. But a careful analysis will show that this cannot be the answer in models based on the Friedmann equation. Causality is the source of this conundrum.

We have described the surface of last scattering of the cosmic background radiation as an impediment to our vision of the universe. However, there is a more fundamental, and more important, limit to what we can see. Since the universe is of finite age, and the speed of light

Why is the universe homogeneous and isotropic?

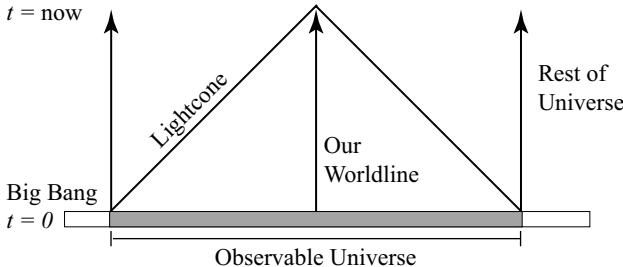


Fig. 16.1 A space-time diagram illustrating the cosmic particle horizon, which defines the observable universe. If we trace our past lightcone back to the big bang, we find the most distant worldline that was ever within our past lightcone. The present distance to this worldline marks the particle horizon limit.

is finite, we cannot see all of it at any specific cosmic time. A surface beyond which we cannot see is called a **horizon**. We have already seen an example of an **event horizon** at the Schwarzschild radius of a black hole. An event horizon is a lightlike (null) surface in space-time that forms the dividing line between those events we can see, and those we cannot. The Schwarzschild horizon occurs when gravity is so strong, as it is around a black hole, that light is trapped; light from inside the event horizon will never reach an outside observer. Another example of an event horizon is the past lightcone itself; it separates the universe into a region that we can see, and regions we cannot. Virtually all the information we can obtain about the universe is carried to us, either directly or indirectly, by light; therefore, when we look into the cosmos, we are looking along our past lightcone. We cannot see events that are spacelike separated from us; at any given moment, there are events that are invisible because they lie outside the lightcone, in our elsewhere.

There is another type of horizon that is of special importance to cosmology. If the universe is not infinitely old, there may exist objects whose light has not yet had time to reach us over the age of the universe. That is, given a time t_0 since the big bang, there exists some distance r_{\max} beyond which all objects are invisible because their light has not yet arrived. Over all directions, this distance would demarcate a two-dimensional sphere in our three-dimensional space. A surface such as this sphere is called a **particle horizon**. The particle horizon represents the tracing of the instantaneous lightcone all the way back to time $t = 0$. Any object whose worldline lies entirely outside this past lightcone is beyond our particle horizon and cannot affect us. More generally, any two objects in the universe can be causally connected, and thus able to influence one another, *only* if they are within one another's particle horizon. The particle horizon changes with time; as the universe ages, more and more objects become visible to us.

In cosmological contexts, the terms “Hubble length” and “horizon” are often used interchangeably. For many applications, such as matter perturbations during the epoch of galaxy formation, there is little practical difference between the two, and even astronomers often do not need to make a careful distinction. However, when we consider the very early universe, we must be more precise. To understand this, first consider the usual lightcone of special relativity. In special relativity, light follows straight lightlike lines in Minkowskian space-time. We can always set

Horizons in the universe

coordinates for our convenience, so let us consider light traveling only along radial directions. The lightlike lines are defined by the condition that the space-time interval is always zero, that is, $\Delta s^2 = 0$. The usual special-relativistic metric equation becomes

$$c^2 \Delta t^2 - \Delta r^2 = 0. \quad (16.1)$$

The lightcone is then

$$r = \pm ct. \quad (16.2)$$

The plus/minus notation merely indicates that the light can travel in either direction. This equation tells us that over a time t , light has traveled a distance r to reach us. We have used this equation, which is strictly valid only for straight lightcones, when we defined the Hubble length.

The lightcones in general relativity are curved by the effects of gravity, so we must use the appropriate general-relativistic metric to compute them; conceptually, however, they are exactly analogous to the lightcones of special relativity. Let us repeat our calculation, this time using the Robertson–Walker metric. We shall again consider only light traveling along a radial path to reach us, so we set the angular change to zero. The light follows a lightlike geodesic specified by

$$c^2 \Delta t^2 - R^2(t) \frac{\Delta r^2}{(1 - kr^2)} = 0. \quad (16.3)$$

The lookback distance is the distance traveled by the light through three-dimensional space over some given time interval. The horizon distance is simply the lookback distance to the beginning of the universe. Because lightlike paths are curved in this space-time, not straight as in the Minkowski metric, we must integrate (sum) the small intervals in both space and time, in order to find the total distance that a light ray could have traveled from the big bang to the present. In this general-relativistic equation for the lightcone, the scale factor accounts for the expansion, and the curvature constant is present due to possible curvature in three-dimensional space. In order to evaluate the required integral, we must know the scale factor R as a function of t ; that is, we must have a model of the universe. It turns out that this sum can be easily calculated for flat space ($k = 0$) with any scale factor that is of the form $R(t) = t^n$. As long as $n \neq 1$, the result is

$$r_h = \frac{ct_0}{1 - n}. \quad (16.4)$$

For $n < 1$ this number never becomes infinite, no matter what the age of the universe; there is always a particle horizon from the very first instant. Because t increases faster than R so long as $n < 1$, the particle horizon always increases, encompassing more and more of the universe with time.

The simplest solution for the scale factor for a standard model is the matter dominated, flat, Einstein–de Sitter universe; recall that $R(t) \propto$

$t^{2/3}$ for this case. Therefore, the horizon length for this universe is

$$r_h = 3ct_0 = 3c \frac{2}{3H} = \frac{2c}{H}. \quad (16.5)$$

This shows that for the Einstein–de Sitter model, the particle horizon length is actually twice the Hubble length. Using this formula with a Hubble constant of $72 \text{ km s}^{-1} \text{ Mpc}^{-1}$, we obtain a horizon distance of about 8×10^9 parsecs for this example. The flat, radiation dominated solution is characterized by $n = 1/2$, which gives a different result for the particle horizon distance.

The particle horizon is the farthest distance to anything whose worldline was ever within our past lightcone; it contains all of the universe that is in any way observable to us. Not only does it limit our view of the universe, but its consequences for causality create one of the most important unsolved puzzles of cosmology. At any given time, parts of the universe are not in causal contact with other regions. The universe became matter dominated within roughly 10^{12} s after the big bang, and this is approximately the time when structure formation began. At that time, only objects that were within a distance of about 10^5 lightyears could have influenced one another. How far apart would two objects separated by this maximum distance be today? Their distance would have increased by the scale of $R(t)$. The redshift at the beginning of the matter era is close to 3000. Hence these two objects are now separated by about 3×10^8 lightyears, or about 100 Mpc!

Around the time of last scattering, at a time of about 10^{13} seconds, the horizon length was on the order of 10^6 lightyears. Across the enormous distance from which we observe it, in both time and in space, the horizon length as of the time of recombination subtends but a small angle on the sky. Expressing this distance as an angular separation of two points on the sky, we can compute that regions now separated by a bit more than approximately 1° could *not* have been in causal contact with one another around the time of recombination. For comparison, the width of the full Moon covers an angular separation of about half a degree; and the full Moon occupies very little of the area of the sky. In fact, this causality scale is what sets the fundamental scale seen in the *WMAP* data. Areas of the sky separated by more than 1° were never in causal contact in the standard big bang model, yet when we measure the temperature of the CBR we find that the deviations from perfect uniformity over the entire sky are tiny. We see this amazing isotropy despite the fact that at the time of recombination, today’s observable universe consisted of approximately a *million* causally *disconnected* volumes!

The only processes we know by which radiation could be forced to the same temperature over a large region require communication throughout the region, and that could never have happened within the standard big bang model. How could conditions have been so phenomenally uniform in all these isolated patches, that the temperature was everywhere the same at recombination despite the fact that the separate volumes of the universe had never been in contact with each other? It is as if there were

Causality in the expanding universe

How can the CBR have the same temperature everywhere?

an orchestra scattered throughout a large stadium, without a director, without music, the musicians never having met previously or having had any prior discussions, yet they spontaneously began to play a piece both in tune, and in tempo. This causality riddle, namely the lack of any physical explanation for the large-scale smoothness of the universe, is often called the **horizon problem**.

The flatness problem

Why is the universe flat?

A number of independent observations now indicate that the geometry of the universe is flat or nearly so. The *WMAP* observations of the fundamental scale of temperature fluctuations in the CBR provide a direct test of the curvature of space. Indirect tests include measurements of the matter density of the universe. The abundances of primordial elements, together with the hot big bang model, indicate that the density in baryons contributes about $\Omega_b \simeq 0.04$, while dynamical methods suggest the total matter density is $\Omega_M \simeq 0.3$. Although some uncertainty remains in the determination of the geometry of the universe, $\Omega = \Omega_M + \Omega_\Lambda$ is certainly close to one. A value for Ω of unity represents something special; in the standard models it is the boundary between an open and a closed universe. Why should Ω be so close to this special value? Why is it not well separated from the dividing line, perhaps $\Omega \simeq 10^{-6}$? Flatness is an unusual condition, lying precisely between spherical and hyperbolic geometries of many different possible curvatures. Among the entire range of possible models that satisfy the Friedmann equation, why is the universe flat or very nearly flat? Might there exist some mechanism that would *require* flatness?

This issue is not merely a persnickety detail. The state of “nearly flat” is much more unusual than it might at first glance appear to be. If Ω is only close to, but is not precisely equal to unity at the present time, at very early times it would have had to approach unity to a fantastic degree. In a closed or open universe, that is, one in which $\Omega \neq 1$, Ω changes with time. For simplicity let us set $\Lambda = 0$ in this discussion. At any particular time Ω is given by

$$\Omega = \frac{\rho}{\rho_c} = 1 + \frac{kc^2}{H^2 R^2}. \quad (16.6)$$

This demonstrates that in a curved space, Ω evolves like $1/(H^2 R^2)$. In the special case of the flat universe, the time dependence vanishes and $\Omega = 1$ throughout history. For a universe of constant negative curvature (Ω less than unity), however, Ω decreases with increasing time, since it heads toward zero as time goes to infinity in an open universe. Therefore, in such a universe Ω was closer to one in the past, and *extremely* close to one during the first few seconds of the big bang. Indeed, for the present universe with Ω near 1 today, at a cosmic time of one second Ω would have differed from unity by roughly one part in 10^{16} , that is, $\Omega(t = 1 \text{ sec}) \sim 1 - 10^{-16}$! This implies that if Ω had been already small early in the big bang, for instance, $\Omega \sim 0.1$, then the universe would be

obviously open, and nearly empty, today. A similar argument can be made for the closed spherical model. In order for Ω to be just barely larger than 1 now, it would necessarily have been infinitesimally above unity at very early times.

As it happens, if the universe were not very nearly flat, we most likely would not exist. If the value of Ω had been much above unity early in the history of the big bang, the universe would recollapse very quickly; it could have collapsed even before galaxies were able to form, much less before life would have time to develop. If, on the other hand, the universe had been strongly open, it would have expanded so rapidly that, over a moderate interval of time, its average density would have dropped too low for structures to form. The resulting universe would be devoid of galaxies, stars, and, presumably, life. Once again, out of all possible universes, very special initial conditions seem to be required in order to create a universe capable of containing beings that could ask questions about it. This geometric specialness of the universe is usually called the **flatness problem**.

Unless the universe is flat, geometry quickly dominates its evolution

The structure problem

If the universe is isotropic and homogeneous on the largest scales, how could any structures exist? In order to produce the structures we observe, the universe must be mostly, yet not completely, homogeneous; significant, though not excessively large, seed perturbations must have existed from which structure could develop. If we assume the existence of the appropriate seed perturbations, we can develop a theory, such as cold dark matter, which at least potentially explains the subsequent formation of structures like galaxies and galaxy clusters. But how did the seed perturbations originate? How could any physical process, which depends upon communication from one region to another, have created in noncommunicating regions coherent inhomogeneities that could have later collapsed? Did the initial conditions somehow collude to create just the right perturbations?

How do similar seed perturbations arise in causally disconnected regions of the early universe?

Even more puzzling, how does the universe create seed perturbations that were approximately the same over all of the observable universe? When we look at opposite points in the sky, we see galaxies that look similar to one another, and to all other galaxies everywhere else that we can see. The mix of galaxy types is not too different around the sky; the clusters also look similar. As far as we can tell, even the great voids and filaments of galaxies are found throughout the visible universe. Why do we not see spirals in one part of the sky, and only ellipticals in another? Or why not galaxies in one region, and elsewhere, in another region that did not communicate with the first at the time when the matter perturbations began to grow, something else entirely? What told all those galaxies to form in a similar way everywhere, even though in the concordance model, regions of the sky separated by more than about twice the width of the full Moon could never have been in causal

contact? Once again, we are stymied in our attempts at understanding by the small size of the horizon at early cosmic times.

The improbability of forming such similar structures in an otherwise isotropic, causally disconnected universe may be called the **structure problem**. It is not a completely independent issue, but is tightly connected to the horizon problem.

The relic problem

Relic particles from the GUT epoch are not observed

A different issue, related to matter density, is what might be called the **relic problem**. Grand unified theories and other advanced theories of particle physics predict a proliferation of massive particles. In some cases, such particles might provide just what is required to explain the dark matter. Generically, however, they tend to be too massive, causing the universe not only to be closed, which does not seem to be borne out by observations, but to have already collapsed, which is definitely not supported by observations!

One such troublesome particle that arises in advanced theories of particle physics is a beast called the *magnetic monopole*. A monopole, in general, is a particle that produces a field of a particular form; for example, any charged point particle such as an electron or proton acts as an electric monopole. In classical electromagnetic theory, there exist electric monopoles (ordinary charged particles) but no magnetic monopoles. Most GUTs predict magnetic monopoles, somewhat to their embarrassment, because magnetic monopoles would have significant and observable effects on the cosmos, none of which have been seen. They would profoundly affect stellar evolution, for example, in ways that might not be pleasant for those of us who depend on one particular, stable star. Since magnetic monopoles, if they exist at all, could not be present in any substantial density in the universe, some mechanism must dilute them after they form during the unified epoch.

The magnetic monopole is just the prime example of an undesirable relic particle; there are others. If such relics exist, they would have been created too copiously in the early universe. No unequivocal detections of particles such as magnetic monopoles have ever occurred, and this places severe limits upon their possible density, at least in the neighborhood of the solar system. The simplest GUT theories predict quite a large density of magnetic monopoles in the present universe. Within the standard models, with no way to eliminate these stable particles once they have formed, they would still be traveling the cosmos, wreaking their havoc; yet we do not observe any trace of their presence. Perhaps it is possible to circumvent this apparent contradiction with various alterations to the simple GUT theories. However, it has proven difficult to construct a GUT theory that does not predict a significant density of monopoles.

The possibility that magnetic monopoles might play an important role in the universe was first suggested in 1931 by P. A. M. Dirac. Dirac argued that the existence of a single magnetic monopole in the entire

universe could account for the discreteness of electric charge. Of course, this was long before the formulation of modern field theories, and Dirac's monopole is not really the same creature as a GUT monopole; still, the concepts have similarities. Magnetic monopoles would also make classical electromagnetic theory perfectly symmetric between its electric and magnetic parts, a symmetry that it now lacks because it contains electric monopoles but no corresponding magnetic monopoles. From various arguments, then, it seems that a magnetic monopole ought to exist. That it apparently does not, at least not in appreciable densities, must be telling us something.

The relic problem is a difficulty that arises not within the standard models of cosmology alone, but rather stems from efforts to combine the big bang cosmological model with particle physics. Nevertheless, it is something that cannot quite be made to fit, and therefore invites an explanation.

A new explanation

The simplicity, elegance, and predictive power of the big bang model are rare qualities in cosmology. Yet we have also seen that it provides no answers to some significant questions, most of which can be reduced to the issue of initial conditions. If our universe is strictly the result of steady expansion from a big bang, then it is a remarkably, perhaps extraordinarily, special universe. Perhaps that is just the way things are, and the universe is an exceptionally low-probability special case. But we cannot draw such a conclusion on the basis of what we now know. It is possible that the physics of the Planck epoch, of which we have almost no understanding, requires that the initial conditions be the way they were. In such a case “the dice were loaded,” and a universe like we see around us was the necessary outcome.

How improbable is our universe?

Let us consider a mundane analogy. Suppose you receive a letter informing you that you have just won a lottery. You know nothing about the conditions under which the lottery was held. You do not know the odds of any particular winning combination; indeed, you do not even know that all combinations are equally likely in that particular lottery. Your only firm knowledge is that a winner, you, exists. Should you be surprised? If the odds are like those of a typical state lottery, then a highly improbable event has occurred. On the other hand, if the letter informing you of your good fortune bears the stamp “bulk rate,” then your win may not be so unlikely after all. We know that state lottery winners do exist. Improbability is not the same as impossibility. But in the case of the universe, we do not know whether our universe really is so improbable. Although we know that the initial conditions seem to have been very special, we cannot yet say for certain that all initial conditions are equally likely.

Perhaps we can answer some of these questions, without recourse to an anthropocentric appeal to the strong anthropic principle, or to vague

Inflation is a de Sitter phase in the early universe

statements about the as-yet-impenetrable Planck epoch. This new explanation goes by the name of **inflation**, or the *inflationary cosmology*. This is not a completely new model, but a supplement to the earliest moments of the standard model. In the inflationary model, the universe undergoes a brief period of rapid exponential expansion early in its history. We have already encountered one solution that inflates: the de Sitter cosmology. This model contains no matter, but only a constant, repulsive cosmological constant. We know that this cannot be a valid model of our current universe, since matter certainly exists. How might it yet provide a reasonable description of an interval of the early universe?

Recall that Λ can be regarded as a **vacuum energy** density.¹ From the perspective of particle physics, a vacuum energy should arise from some quantum field, presumably a field involved in gravity. Quantum fields are subject to the uncertainty principle. A consequence of the application of the uncertainty principle to quantum fields is that there is no such thing as a constant field. All fields undergo continual changes at the level allowed by the uncertainty principle. These random and unobservable changes are called **quantum fluctuations**. Although quantum fluctuations themselves might be unobservable, they can have physical, observable effects. An example of such a phenomenon is the *Casimir effect*. Imagine two perfectly clean metal plates in a perfect vacuum, separated by a very tiny gap. If the plates are electrically neutral, then classically there should be no force between them. However, the presence of the plates restricts the wavelengths of the quantum fluctuations permitted to exist between them, while the vacuum outside the plates has no such restriction. Thus there are fewer virtual particle pairs between the plates than are present in the region surrounding them;² the vacuum between the plates is “more empty” than that outside. The consequence of this effect is a very small attractive force pushing the plates toward one another, as if there were a *negative* energy density, or negative pressure, in the region between the plates. This may sound like science fiction, but recent technology has enabled physicists to measure the Casimir force to very high accuracy, and it was found to obey the theory exactly. Quantum fluctuations are a genuine aspect of nature.

Unfortunately, while quantum mechanics tells us how a vacuum energy density can arise, it so far has failed to permit us to calculate the value of Λ . Attempts to use considerations of spontaneous symmetry breaking to obtain a naive estimate of the vacuum energy density would predict a value comparable to the fourth power of the *Planck energy*, which is the scale at which gravity is thought to have decoupled from the other forces. The Planck energy represents a fundamental scale of physics, since it is that energy obtained by combining the fundamental constants of quantum mechanics (Planck’s constant h), gravity (G), and electromagnetism (c) to create a quantity with units of energy. Specifi-

¹See Chapter 13.

²See Chapter 9 for a discussion of virtual particles.

The Planck energy

Regular Vacuum – All Quantum Wavelengths Allowed

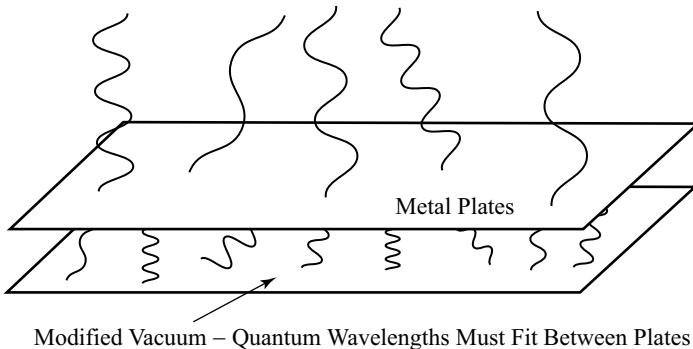


Fig. 16.2 The Casimir effect. Two metal plates in a vacuum restrict the wavelengths of the quantum fluctuations that are possible. Outside the plates the vacuum has no such restriction. The difference creates a slight attractive force, pushing the plates together.

cally, it is

$$E_P = c^2 \left(\frac{hc}{2\pi G} \right)^{1/2}. \quad (16.7)$$

Working this out in MKS units, we find that the Planck energy is approximately 2×10^9 joules. In units that might be more familiar, this corresponds to about 550 kilowatt-hours, or roughly half a ton of TNT equivalent. This is an enormous energy on the scale of any elementary particle; in comparison, the rest energy of the proton is only 1.5×10^{-10} joules. When the Planck energy is assigned to the quantum of a field (that is, a particle), and the corresponding energy density of the field is computed, the fourth power of the resulting number is at least 120 orders of magnitude greater than the present estimate of the vacuum energy density. This mismatch between the only known natural scale for the cosmological constant and the observational limits on its value is not understood. The smallness of the present-day vacuum energy density is a conundrum similar to the flatness problem. The vacuum energy density is so tiny compared to the Planck energy that it “ought” to be zero. There is no physical explanation for why it should be some relatively small, but nonzero, value, rather than a function of the fourth power of the Planck energy, or the energy associated with some other symmetry breaking.

Thus the inflation model must hypothesize that a quantum field associated with some particle not present in the universe today created the required vacuum energy associated with a nonzero Λ . How can we know what that particle is, or anything about its properties? No particle accelerator we can imagine could reach the energies at which such a mysterious particle might be detected. We must rely upon theory until better data are at hand. Grand unified theories, of which there are several varieties, predict the existence of massive particles that we cannot yet detect. These hypothetical particles play various roles within the theories; which particles are required, and of what type and mass, depends upon the particular theory.

Some candidates exist for providing a vacuum energy density in the early universe. The hypothetical *Higgs boson* is a massive particle that

A hypothetical quantum field

plays a role in many theories of spontaneous symmetry breaking during the unified epoch. The Higgs boson takes part in particle interactions connected to baryogenesis, then conveniently annihilates after the unified epoch; thus it would play no role in the present universe. A more troublesome grand-unified particle is the aforementioned magnetic monopole, which tends to be overproduced by GUTs. These two particles provided the original motivation for inflation. Alan Guth realized in the early 1980s that if the universe did undergo a de Sitter phase during the unified epoch, then the density of magnetic monopoles, which would have formed prior to this era but not during or afterward, would have been diluted away to almost nothing by the huge increase in the volume. All that was needed was a means to induce exponential expansion. The original models of Guth and others invoked the Higgs boson to fill this role. Like all particles, the Higgs boson is associated with a field. The field in turn is associated with a **potential**, a function in space-time that describes the energy density of the field. From the potential, we can learn where and when the value of the vacuum energy density might have been nonzero.

These concepts are awkward to express in words; here we see an example of how mathematics tremendously increases our power to understand quantitative ideas. It is far more transparent to use elementary mathematical notation when we work with the ideas behind inflation. Suppose that the field, which is just a description of a variation in space and time, is specified by some function $\phi(x, t)$. The quantity x stands for any one spatial dimension; examples with a single spatial dimension suffice for purposes of illustration. We do not need to know any details of the field here. All we assume is that the field is *scalar*; that is, it has a single value at each point, so that we do not require multiple components to specify the behavior of the function. All advanced theories of inflation start from a scalar field, so this is not an excessive simplification. Now we suppose that the field has an associated potential $V(\phi)$. That is, V is a function of the field quantity ϕ , and thus V depends upon space and time only indirectly. If we wish to know the potential at any event (x, t) , we first compute ϕ at that event, then insert that value into the potential function $V(\phi)$. As a concrete example, we could consider the height above sea level as depicted on a topographic map to be a scalar field, and the gravitational potential to be the associated potential function. In this case, the gravitational potential tells how much energy is acquired or expended in moving from one height to another. In the case of the early universe, computation of the potential gives the energy density at that event.

Consider some particular event (x_i, t_i) . The field, and thus the potential, have certain values at that event, specified by their respective functions. Now contemplate the region in space-time near this event. If the potential is greater for those values of x and t that are a little different from x_i and t_i , then the field will prefer to sit where it is. If the potential is smaller in some direction, the field will tend to shift toward that direction. The general principle is that the field attempts

A scalar field and its potential

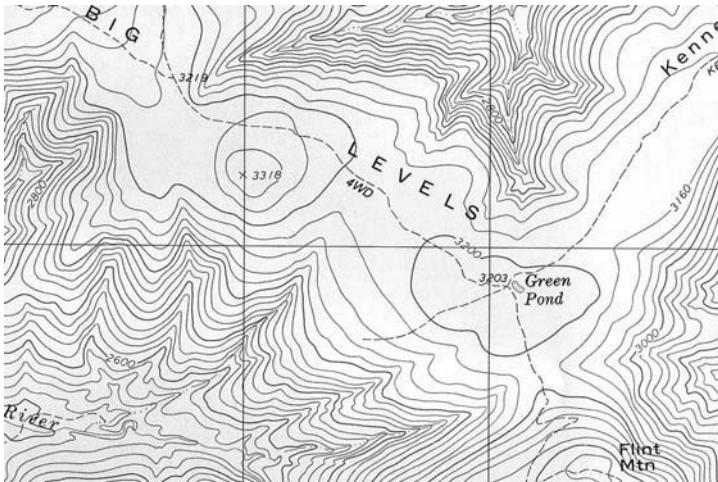


Fig. 16.3 A contour map shows altitude above sea level. Each point has an elevation; this is a scalar function of position $\phi(x)$. The elevation in turn implies a gravitational potential $V(\phi)$. It is energetically favorable for the mountains to erode into flat plains, that is, to evolve to a uniform minimum altitude with lower potential.

to minimize its potential. The analogy of potential to altitude is once again helpful. In the gravitational case, it is energetically favorable for mountains, which represent maxima in altitude, to erode away and form a smooth flat plain, since this would minimize the gravitational potential of the terrain. Similarly, if the region around (x_i, t_i) is very flat, as defined by the form of $V(\phi)$, then the field will not change; that is, we say it is *stable* around this event. For example, it may happen that for a certain time interval, which may be short or long depending upon the potential, the spatial variations of the energy density are very small. The field will remain stable until, perhaps after some appropriate time interval has passed, the form of the potential function changes. If the potential function changes, the field will have to adjust anew.

Now let us apply these general ideas to a scalar field in the early universe, and return to the hypothetical Higgs boson as an example of a particle that is associated with a scalar field. The Higgs boson controls a symmetry breaking in the unified epoch. When the mean energy per quantum of the universe is above the scale set by the breaking of this particular symmetry, it turns out that the Higgs field is stable, with an expected, or average, value of zero, in the vicinity of some positive value of the potential, that is, $V(\phi = 0) = V_0 \neq 0$. Because this energy density is greater than zero when its allied field is zero, this state is known as the **false vacuum**. It is not an absolutely stable state, but is like the stability of a marble resting atop an inverted bowl. The marble will stay there as long as it is not disturbed, but a perturbation will cause it to roll down to the genuinely stable state on the table. The potential of the Higgs boson resembles such an inverted bowl with a very flat base. The false vacuum is a *metastable* state; as long as no perturbation occurs, the field will remain in that state, analogous to the marble sitting on the base of the bowl.

While the field is in the metastable false vacuum, the vacuum energy density dominates the universe completely. As we have discussed, it acts

An example of inflation

like a negative pressure, a large one in this case, which can be shown to obey an equation of state

$$P = -\rho_\Lambda. \quad (16.8)$$

The energy density in a given volume of vacuum is constant, hence ρ_Λ is constant. Inserting a constant density into the Friedmann equation (11.10) yields a term that varies as R^2 , exactly as a cosmological constant does. Comparing the constant-density form of the standard Friedmann equation with the way in which the cosmological constant Λ enters the generalized Friedmann equation (11.22) shows that the vacuum energy density can be equated to a cosmological constant:

$$\rho_\Lambda = \frac{\Lambda}{8\pi G}. \quad (16.9)$$

So far this would give us only the familiar de Sitter solution, which has nothing to say about the physical origin of the cosmological constant. For a flat geometry, we have seen that the de Sitter model expands exponentially; hence the inflation. Any positive cosmological constant will eventually cause the universe to inflate, regardless of its source.

If we assume that the cosmological constant is a consequence of the energy density of some quantum scalar field, then we must make a connection between the potential and the cosmological constant. This relationship would vary from one field theory to another, but for the sort of theory we are taking as our example, namely, a field connected to a spontaneous symmetry breaking, we can consider the constant vacuum energy density to be specified by the value of the potential function when the field is zero. Symbolically,

$$\rho_\Lambda = V_0. \quad (16.10)$$

For most theories of this class, it happens that $V_0 \sim E^4$, where E is an energy that characterizes the scale of the symmetry breaking. This may provide some justification, though not an explanation, for our earlier assertion that the natural scale of the cosmological constant should be something like the fourth power of the Planck energy. For the particular field we have used as an example, the Higgs field, the energy in question would *not* be the Planck energy, of course; it would be the rest energy of the Higgs boson, which is still very large because the strong and electroweak interactions are unified only at very high energies. If the Higgs boson had been responsible for an inflationary phase during the unified epoch in the early universe, then the cosmological constant during that interval would have been given by $\Lambda \sim 8\pi G E_{\text{Higgs}}^4$. The energy density of the false vacuum, acting like a cosmological constant in a de Sitter universe, powers the inflation. Once the energy falls below the scale of unification, the field adjusts and drops down to a state of zero vacuum energy density, the **true vacuum**, and inflation ceases. The true vacuum is genuinely stable, and the field remains there indefinitely.

The Higgs boson is not established, however. It is an example of a particle that does play a well-defined role in a theory of grand unification, and which does have some characteristics that would have to be

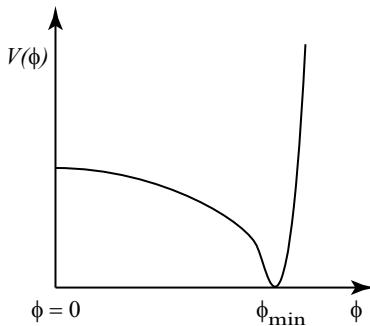


Fig. 16.4 Schematic illustration of an inflationary potential. The false vacuum occurs because $V(\phi)$, which represents the vacuum energy, is nonzero when the field ϕ is zero. The field eventually transitions to the true vacuum of zero potential, for some associated scale given by $\phi = \phi_{\min}$. The universe contains a large energy density due to the field during this transition, and inflation occurs.

present in a particle capable of causing inflation. Historically, it was the particle that motivated the inflationary model in the first place. But it turned out that the Higgs boson is probably not the responsible particle. The original model of inflation described above suffered in the details. Inflation occurred while the field was in the metastable false vacuum. However, for various technical reasons, nothing subsequently could eject the field from that metastable state; the inflation never stopped. This clearly contradicts our observations of the universe, at least that which we can see, so the original version of the inflationary universe cannot be correct. The model was rescued by a change in the potential V . In this new inflationary scenario, inflation occurred not while the field was in the false vacuum, but during the transition from the false to the true vacuum. Since the potential plays the role of the cosmological constant, we can see that a slow decrease during inflation will provide a simple way for the inflation to come to an end after a period of time. In other words, the cosmological “constant” changes slowly during the inflation. To ensure that enough inflation occurs, the potential must be extremely flat, so that the field carries out the transition very slowly; slowly, that is, in comparison to the characteristic rate of expansion at that time.

If such an inflation occurred, it would have happened around 10^{-37} seconds after the big bang, and required approximately 10^{-32} seconds to complete. During this cosmic eyeblink, the scale factor would have inflated by a dizzying factor of some 10^{40} to 10^{100} or even more. This more successful new inflationary scenario was first proposed by Andrei Linde, and independently by Andreas Albrecht and Paul Steinhardt; it was subsequently developed by many researchers. New inflation does not depend upon the presence of any particular particle but merely requires the existence, at the appropriate phase in the history of the universe, of some particle with an extremely flat potential and a slow transition to the true vacuum. This generic particle has come to be known as the **inflaton**.

Since temperature drops as the inverse of the scale factor, it might be expected that inflation would cool the universe tremendously, and it does. How do we reconcile this with the good agreement between observations and the standard hot big bang model, which never undergoes such a drastic expansion? At the end of the inflationary period, the vacuum energy from the earlier false-vacuum and inflation stage is consumed in the creation of particles, which decay and ultimately convert their energy into reheating the universe. A huge energy density would have been locked up in the quantum field, so once this energy was released and converted into other, more conventional, forms of energy, the universe would naturally be reheated back to where it should have been if inflation had not occurred. Before inflation, the evolution of the universe was dominated by particle physics. After this burst of heating, the mostly classical, radiation dominated phase of the standard models began.

The causative agent of inflation remains unknown

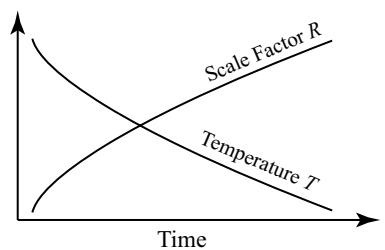


Fig. 16.5 The evolution of the scale factor R and the temperature T in a standard universe. Temperature falls smoothly and is inversely proportional to the scale factor R .

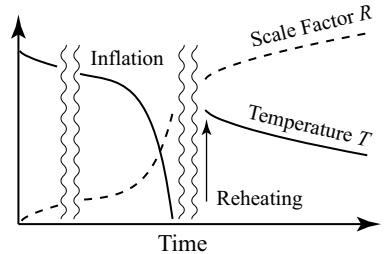


Fig. 16.6 The schematic evolution of the scale factor R (dashed line) and the temperature T in the inflationary universe. (The curves are not to scale.) The exponential expansion of inflation results in strong cooling. At the end of the exponential phase, reheating occurs as the vacuum energy is converted into more conventional energy. The universe subsequently evolves as in the standard model.

Inflation explains isotropy and flatness

Inflation and the initial conditions

The inflationary epoch achieves the goal of explaining the horizon, flatness, and relic problems by rendering the initial conditions insignificant. What we take for special conditions in the universe, such as the observed properties of flatness and isotropy, now arise naturally as necessary outcomes of inflation. For example, the horizon problem is explained by the enormous increase in the scale factor; what is now the observable universe began from a tiny region that *was* in causal contact prior to the beginning of inflation. Thus the isotropy of the cosmic background radiation is automatically assured, and the horizon problem is eliminated. Similarly, the exponential expansion means that the observable universe becomes effectively flat, regardless of its original state, because any curvature it might have possessed initially is stretched away. We can see this by considering the Friedmann equation for a vacuum energy density Λ ,

$$\left(\frac{\dot{R}}{R}\right)^2 = \frac{\Lambda}{3} - \frac{k c^2}{R^2}. \quad (16.11)$$

Since the vacuum energy density Λ is constant, the total vacuum energy increases significantly as the universe expands. The scale factor R grows enormously during this interval and the curvature term drops rapidly toward zero. At the end of inflation this vacuum energy Λ becomes a matter energy density ρ , and the observable universe effectively expands thenceforth as a flat ($k = 0$) universe, regardless of the original value of k .

The effect upon the spatial curvature can be visualized by imagining the surface of a balloon blown up far beyond its normal proportions, until the surface is stretched so much that any small neighborhood around a point will appear flat. Such a simple analogy should not be taken excessively literally, of course. The inflated universe is stretched by such an unimaginably huge factor that any relict curvature has a scale that is enormously greater than the Hubble length. There is no flatness problem in an inflationary universe, since the observable universe would always end up indistinguishable from a flat universe. The critical density is not a special condition in this model, but is the inevitable consequence of inflationary physics in the early universe.

Inflation explains the lack of relic particles

The relic problem is solved because any relic particles created before inflation, such as magnetic monopoles, are exponentially diluted away, their density diminishing to insignificance. This means that baryogenesis must occur *after* inflation, or else the baryon density would also have dropped effectively to zero and galaxies as we see them would never have formed. However, after the universe reheats, baryogenesis proceeds exactly as in the standard model, so this is not a problem for inflationary scenarios.

In solving these problems, have we aggravated the structure problem? If the universe now can stretch away any inhomogeneities, where are the seeds for the formation of structure? It was quickly realized that infla-

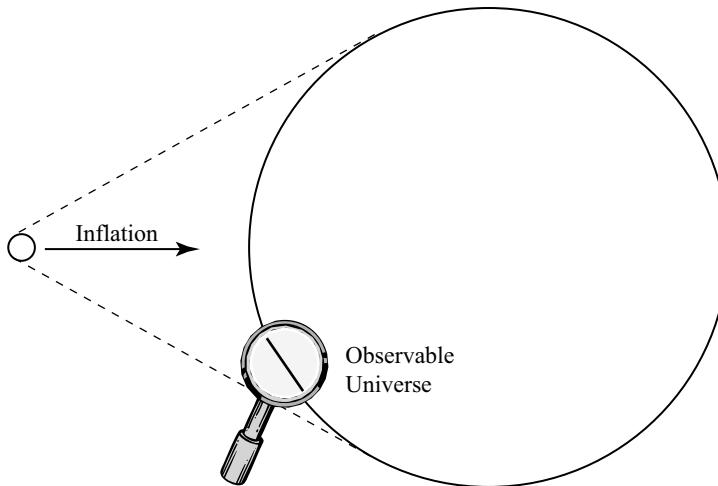


Fig. 16.7 A small universe undergoes exponential inflation by a huge factor. Any initial curvature is enormously stretched. Our observable universe, a tiny bit of this inflated structure (here magnified for view) has no measurable curvature within the Hubble length. Thus, in the inflation picture it is natural that we observe a flat universe.

tion itself provides an explanation for the seed perturbations required for structure formation. According to the inflationary scenario, these perturbations originate in the quantum fluctuations of a field associated with some kind of elementary particle. We have seen such a process before in the discussion of Hawking radiation around black holes.³ In Hawking radiation a pair of virtual particles is created near the horizon; one is trapped within the horizon and the other, no longer able to annihilate with its counterpart, becomes real and escapes as outward-traveling radiation. In the inflationary universe the inflation is so rapid that virtual particles are pulled apart and become causally disconnected, thereby becoming real, rather than virtual, fluctuations in the field. Like other lengths, the sizes of these fluctuations are stretched by the enormous increase in the scale factor. They become larger than the horizon size of the universe and are frozen into the background. These frozen fluctuations are amplified by the inflation and become large enough to generate macroscopic inhomogeneities once the universe reverts to a standard, noninflationary evolution. It turns out that this process naturally leads to the Harrison–Zel'dovich spectrum of perturbations; that is, a power spectrum with equal power on all scales, the equivalent of acoustical white noise. This is a very pleasing result; this spectrum was proposed long before inflationary models were developed because it had properties that seemed to fit the observed structures well, and because it required minimal assumptions about the nature of the perturbations. Its appearance in a well-motivated model provides a physical explanation for what had previously been justified primarily on mathematical and empirical grounds.

The question of the origin of the seed perturbations may seem to be a somewhat mundane technical issue, but consider the implications of the answer that inflation provides. If we think of quantum mechanics at all,

The origin of structure within the inflationary universe

³See Chapter 9.

we are used to relegating its properties to the smallest imaginable scales in the strange world of subatomic particles. Yet inflationary cosmology proposes that the *largest* scales in the universe, the great galactic super-clusters, the filaments and sheets of structure, and the voids, are nothing but quantum fluctuations writ large. The largest things in the universe mirror the structure of the shifting nature of reality at the quantum level.

Chaotic inflation

Larger implications of the inflationary model

Inflation wipes away the memory of any initial conditions in the big bang singularity and ensures that the observable universe will be flat and smooth, regardless of how it began. Inflation also naturally provides the necessary quantum seed perturbations from which to form structure. But if our entire observable universe began as a tiny patch within a larger big bang, does that mean that our universe is not all that exists? Are there other universes? Inflation will occur only in patches where the vacuum energy of the false vacuum is dominant; this condition is not guaranteed to hold over the entire universe during the unified epoch. The inflating field would not likely be constant everywhere. The most general case is **chaotic inflation**, an idea originally proposed by Andrei Linde.

In the chaotic inflation model, quantum fluctuations in a primordial field cause some portions of the universe to inflate considerably, while others expand to a lesser degree. In some regions, the inflation falters altogether. In this scenario, the inflating patches rapidly form child universes that are attached by wormholes to the mother universe. The wormhole effectively cuts off communication between the child and mother universes. Our observable universe might be only one of many noncommunicating universes, some of which have inflated, and appear flat, and others of which retain the original initial curvature! Presumably this process took place from the indefinite past and continues into the indefinite future.

An interpretation of the anthropic principle

Chaotic inflation offers an intriguing interpretation of the anthropic principle. The quandary posed by the anthropic principle is that many things in the universe seem arbitrary, such as the mass of the proton, the relative strengths of fundamental forces, and the speed of light. Yet it seems that if these constants of nature had not taken the value that they have, then galaxies, stars, planets, and life would not have come into existence. If this were truly the one and only universe, the sum total of all there is, this would seem to be a remarkable thing. In the chaotic inflation scenario, however, it is not particularly significant that conditions seem to be remarkably right in our universe for life to have formed. Each child universe in the chaotic inflation model could well have its own set of physical conditions. When there are many universes from which to choose, only those in which the correct conditions prevailed would have given rise to life forms who could ask questions about

their environment. In such a case, marveling at the apparent wonder of our existence is like speculating about the astonishingly unlikely combination of genes that any human possesses. The specific combination of genes that makes up any given individual is almost infinitely improbable, from a strictly statistical point of view; yet no one is surprised about the existence of any particular person. This idea of multiple, possibly infinite, child universes seems like a natural way out of the “specialness” problem. However, the special conditions of our universe do not by themselves constitute a proof of the chaotic-inflation model.

We can compare this to a game of straight poker, in which each player is dealt five cards and must play them as they fall. For our purposes, let us define a hand as some combination of five cards from a deck of 52. The number of such hands is enormous, more than two and a half million. The probability of receiving *any* hand of five specific cards is equal, approximately one in two and a half million; but sometimes there is more than one way to achieve a certain scoring combination. For example, there are many ways in which a pair, or even three of a kind, could occur in a hand, but only four royal flushes (ten, jack, queen, king, and ace of one suite) are possible; hence the probability of a royal flush is roughly four in two and a half million. The *gambler’s fallacy* is the assumption that a long sequence of poor hands means that a good hand is somehow due; most of us know, at least rationally, that this is not the case. Now turn this around; suppose that the first thing a visitor to a poker game sees is a player laying down a royal flush. No doubt he would be amazed. If he then concludes that a large number of hands must have been played previously, he would have fallen into the *inverse gambler’s fallacy*. The occurrence of an improbable event does not imply that any previous trials need have taken place at all. For instance, the winner of a state lottery may have never played before. Thus the apparent specialness of our universe does not in itself *require* that other universes exist; this would be the inverse gambler’s fallacy. In cosmology, unlike poker, we do not even know yet what the odds are.

The chaotic inflationary model would be difficult or impossible to test observationally, since the individual child universes cannot communicate. It has been suggested that it might be possible to create miniature quantum bubbles in the laboratory, but this experiment would seem to be extremely difficult (and perhaps dangerous). Indirect tests might be possible, if ever the inflaton is detected, so that its associated potential can be determined and compared with theory. An improved understanding of grand unified physics may show that chaotic inflation is the inevitable outcome of the pre-inflation state. For now, however, chaotic inflation is another interesting speculation on the frontiers of cosmology.

The difficulty of assessing probability without data

Testing inflation

Something as seemingly bizarre as inflation prompts the suspicion that cosmologists have ventured beyond the scientific and testable into the

purely speculative. However, inflation can be put to the test. After all, Guth's original proposal for inflation was found to be unworkable, and hence was falsifiable. Physicists continue to work toward GUTs, quantum gravity, and theories of the high-energy state of the early universe. A direct consequence of such a new theory may very well be something like inflation. And inflation in its present form makes some predictions about the subsequent evolution of the universe that can be tested against observations.

Inflation makes predictions about fluctuations in the CBR

High-resolution *WMAP* data on the temperature anisotropies in the CBR provide a means of testing one fairly universal aspect of nearly all inflationary models. Inflationary models predict a particular spectrum of perturbations arising from quantum fluctuations; namely, random fluctuations with equal power on all scales. Many other proposals for the initial seed perturbations produce different perturbation spectra, so it is possible to distinguish among various hypotheses. The angular size and amplitude of the CBR temperature anisotropies depend upon the nature of the seed perturbations. Of course, they also depend upon many parameters of the universe such as spatial geometry, baryon density, Hubble constant, and the like, so tests of inflation are only part of the overall data analysis effort. From the results so far obtained, *WMAP* confirms the basic ideas of the inflationary picture and provides some preliminary quantitative tests of the details. Experiments and data analysis will continue to provide ever more stringent tests. The launch of the *Planck* satellite will be an important milestone. *Planck* will operate with greater sensitivity and resolution, and should significantly improve the study of the temperature anisotropies.

Another obvious consequence of the standard inflationary model is a perfectly flat universe. This too was confirmed by the observations of *WMAP*. Measurements of Ω_M alone point toward $\Omega \sim 0.3$, implying that the balance must be found in the dark energy, or Ω_Λ . Inflation is not incompatible with the presence of Λ in the universe today. It would, however, be a different Λ from the one that drives inflation in the early universe. Unfortunately, at the present time the existence of Λ , or dark energy, means that we exchange one special situation, a flat universe, for another, a cosmological constant that is nonzero and makes a dynamical contribution to the universe comparable to that of the matter. It is puzzling that at the present time Ω_M is quite close in value to Ω_Λ , but it does seem to be what the current observations demand.

Limitations of inflation

Does inflation solve all our cosmological problems? Unfortunately, it does not; there are still mysteries in cosmology. The model does not explain why the present-day Λ has the nonzero value it apparently has. Inflation also does not yet seem to fit comfortably within any known scheme of particle physics. The Higgs boson was originally thought to be the particle responsible for inflation. However, the isotropy of the CBR sets a limit on the amplitudes of seed perturbations; this condition demands an extremely flat potential, which seems to be itself a special requirement. Further investigation showed that a potential that could be

associated with a Higgs boson was not sufficiently flat to cause inflation without also disrupting the CBR. Several other potentials have been suggested, most of which have some justification in particle physics, but none of which corresponds to a particle that has any role other than to produce inflation. It does not seem fully satisfying to replace one set of *ad hoc* requirements, the special initial conditions, with another, a particle that seems to have nothing else to do with particle physics, at least not according to current understanding.

It is mysteries such as these that drive progress. Work will continue; perhaps some day a theory will arise that will explain all the data within a fully developed combination of particle physics and general relativity. Unfortunately for the curious among us, no such theory is in sight at present, but a breakthrough is always possible. For the time being, inflation must be regarded as a promising and intriguing hypothesis that has passed several observational tests, but as yet cannot provide all the consistent answers we seek. All sciences, including cosmology, must ultimately be founded on empirical evidence. Better data will guide theoretical progress, and from improved theory we will achieve greater understanding of the origins of our universe.

Chapter Summary

Despite its successes, the standard big bang cosmology has some problems that are difficult to resolve: (1) The *horizon problem* asks how the universe became so highly homogeneous and isotropic when most regions of the observable universe were not in mutual causal contact at early times. (2) The *flatness problem* is that the universe is nearly flat today, but if it is not perfectly flat, this implies that the value of Ω must have been very nearly equal to unity at early times. This would seem to require extreme fine tuning if the universe is not exactly flat. (3) The *structure problem* asks what formed the perturbations that lead to the structure we see around us. Why is structure the same everywhere, even though different parts of the universe were not causally connected early in the big bang model? (4) The *relic problem* occurs because grand unified theories (GUTs) predict massive particles that are not observed. What happened to these relics of the unified epoch?

The inflationary model addresses all these issues by presuming that what we call the observable universe is actually a very small portion of the initial universe that underwent a de Sitter phase of exponential expansion around the time of the unified epoch. This model posits that what became our observable universe was small

enough to be in causal contact at the big bang; it then grew exponentially during the inflationary epoch. This enormous growth had the effect of flattening out any curvature, stretching the geometry of the universe so much that it became flat. Any massive GUT particles were diluted, spread out over this now fantastically huge domain to an extremely tiny density, so that they no longer are observable. Quantum fluctuations in the vacuum were preserved and inflated to large scales by the expansion, providing the seeds for structure formation.

The source of this exponential growth was a negative pressure produced by a nonzero vacuum energy. A nonzero vacuum energy could result from quantum processes in the early universe. In quantum field theory, a field is associated with each particle, and the field in turn is related to a potential, the latter being a function that describes the energy density of the field. The appropriate potential would result in a false vacuum, a situation in which the field was zero but the corresponding potential was not zero. The false vacuum state could have provided a vacuum energy that would behave exactly like a positive (repulsive) cosmological constant, resulting in a temporary de Sitter phase during which the associated patch of universe grew by a factor of perhaps 10^{100} or

more. Eventually, however, this vacuum energy was converted into real particles and the field found its way to the true vacuum, bringing the inflation to a halt. The universe then continued to evolve from this point as in the standard model.

The inflationary model has had many successes, but it remains an area of active research. It makes some predictions about the structures in the universe that are consistent with the *WMAP* data, and it predicts that the

present universe should be flat, as is observed. However, the particle that might have provided the vacuum energy density is still unidentified, even theoretically; it is sometimes called the *inflaton* because its sole purpose seems to be to have produced inflation. Despite this, it seems difficult to understand how our homogeneous, isotropic, and flat universe could have developed unless something like inflation occurred.

Key Term Definitions

horizon Any surface that demarcates events that can be seen from those that cannot be seen.

event horizon A lightlike surface that divides spacetime into two regions; that which can be observed, and that which cannot. The Schwarzschild radius of a nonrotating black hole is an event horizon.

particle horizon A surface beyond which we cannot see because the light from farther objects has not had time to reach us over the age of the universe.

horizon problem The conflict between the observed high uniformity of the cosmic background radiation and the fact that regions of the sky separated by an angular size of more than approximately one degree could not have been in causal contact at the time of recombination.

flatness problem The observed fact that the geometry of the universe is very nearly flat, which is a very special condition, with no explanation of why it should be flat.

structure problem The difficulty of explaining the origin of structure, representing local inhomogeneities, in a universe that is isotropic and homogeneous on the largest scales.

relic problem The problem in standard cosmology in which various theories of particle physics would invariably produce massive particles that are not observed.

inflation A period of exponential increase in the scale factor due to a nonzero vacuum energy density. This occurs early in the history of the universe in certain cosmological models.

vacuum energy The energy associated with empty space, that is, the vacuum itself.

quantum fluctuation The small variation that must be present in a quantum field due to the uncertainty principle.

potential In physics, a mathematical function that describes the energy density of a field.

false vacuum A metastable state in which a quantum field is zero, but its corresponding potential is not zero.

true vacuum A stable state in which a quantum field is zero and the corresponding potential is also zero; that is, the vacuum energy density is zero.

inflaton The generic name of the unidentified particle that may be responsible for an episode of inflation in the very early universe.

chaotic inflation A model in which many distinct universes form from different regions of a mother universe, with some inflating and others perhaps not.

Review Questions

- (16.1) Explain the difference between a *particle horizon* (16.5) and an *event horizon*.
- (16.2) (More challenging.) Write down the formula for the horizon length during the radiation era. Compute the horizon length at cosmic time $t = 10^{-37}$ seconds. Consider two particles separated by this distance at this cosmic time. How far apart would they be, assuming only cosmic expansion, at recombination, $t = 10^{13}$ seconds? Assume that the radiation dominated formula for the scale factor still holds. Use meters as your unit for this problem.
- (16.3) Discuss the horizon, flatness, structure, and relic problems. Do they constitute genuine problems for the standard model, in the sense of being inconsistent with its assumptions? In your opinion, which of these problems is most troublesome? Why?
- (16.4) What effect would a negative pressure have had in the early universe? How might such a phenomenon have been produced? What connection might this have to the cosmological constant?
- (16.5) Explain qualitatively what a *potential* is. What is the essential property that a potential must have for inflation to occur? Illustrate your answer with a sketch.
- (16.6) Distinguish the true vacuum and the false vacuum. Which is important for inflation? Why?
- (16.7) How does the inflationary model account for the formation of structure? What features of inflation are particularly appealing in this regard?
- (16.8) If the universe inflated enormously, what would happen to the temperature of any pre-inflation constituents? What is the source of the energy for reheating at the end of inflation?
- (16.9) Discuss how inflation solves the horizon, flatness, and relic problems. What problems does inflation itself introduce?
- (16.10) Discuss chaotic inflation. What intriguing interpretation of the anthropic principle does it offer?
- (16.11) What cosmological observations provide support for the inflationary picture?

This page intentionally left blank

The Edge of Time

17

Not until the empirical resources are exhausted, need we pass on to the dreamy realms of speculation.

Edwin Hubble, *The Realm of the Nebulae*

In our journey through the universe, we have encountered many wonders. Once we leave behind the bounds of our Earthly velocities and distances, we realize that our cozy intuition is often wrong, our common sense does not apply. Special relativity seemed so strange at its introduction that even some of the most distinguished scientists of the day refused to accept it. General relativity had an easier time, but quickly gained such a formidable reputation that most people became convinced that none but the brightest genius could hope to grasp its concepts. Yet both these theories are elegant and straightforward in their fundamental ideas; confusion occurs because they demand a way of thinking that is so at odds with our everyday experience and intuition. However, the special and general theories of relativity are at their cores pure classical physics, the appropriate extensions of Newtonian physics to space-time itself. For most of our story, we have been concerned with astronomically sized objects and with scales, of both time and distance, that are enormous even in comparison to our solar system; therefore, we have easily been able to remain within the realm of classical physics. Yet every now and then, the shadowy world of quantum mechanics has intruded even into our modern version of Newton's clockwork.

Relativity and quantum mechanics were the two great triumphs of 20th century physics. Both were developed during the first 30 years of the century. Both are spectacularly successful within their respective domains. Quantum mechanics governs the world at the smallest scales, the level of particles, atoms, and molecules, while general relativity, as a theory of gravity, rules the largest scales, from stars and planets to that of the universe itself. Low-energy quantum mechanics, as well as special relativity, boast ample experimental verification from the laboratory. General relativity is much more difficult to test experimentally, so its empirical foundation rests upon its success at explaining and predicting certain observed astronomical phenomena. Nevertheless, every such test has produced results completely consistent with the predictions of general relativity.

Key Terms:

- quantum gravity
- Schrödinger equation
- wavefunction
- quantum state
- Copenhagen interpretation
- superposition of states
- collapse of the wavefunction
- measurement problem
- quantum cosmology
- wavefunction of the universe
- entropy
- arrow of time
- grandfather paradox
- many-worlds interpretation
- string theory

Gravity is by far the weakest force in the universe; in the hydrogen atom, the electromagnetic force between the proton and the electron is about 10^{40} times as great as the gravitational force between them. This is fairly typical of the difference in scales between the quantum and gravitational realms, and accounts for our ability, through most of our study of cosmology, to separate the two theories without ambiguity. Yet it is certain that they must inevitably meet. Near a singularity, the curvature of space-time must be so great that the scale of gravity becomes comparable to that of the other forces. To describe such a state, we must find a theory of **quantum gravity**. Moreover, quantum mechanics has already been applied to the explanation of the electromagnetic force and the strong and weak nuclear interactions; should not gravity be similar to the other three fundamental forces? It might seem as though the challenge of developing quantum gravity should not be so great. After all, *special* relativity and quantum mechanics were united in the 1920s by Paul A. M. Dirac. The most significant result of this theory was its requirement that antiparticles exist, a prediction that was confirmed in 1932 by the discovery of the positron.¹ The Dirac theory is now well established as the special-relativistic quantum mechanics. However, general relativity has still not been successfully incorporated into a consistent quantum formulation.

The center of a black hole marks a singularity in space-time, where classical general relativity must break down. We are able to take into account the quantum mechanical nature of matter in white dwarfs and neutron stars because we understand the behavior of matter under the pressures and densities encountered in these objects. For a black hole, we have no such understanding. In the collapse to an infinitely dense singularity, the physics of gravity necessarily enters the quantum realm of the microscopic. We simply do not know how to describe the properties of matter under such extreme conditions, but we know that at some point, quantum mechanics *must* play a role. Without a theory of quantized gravity, we cannot be certain what lurks at the center of a black hole. We do not even need to go all the way to the singularity to find quantum effects associated with black holes; Hawking radiation shows that the very strong gravitational field near the hole's event horizon has predictable quantum consequences. What other phenomena might we discover from a full theory of quantum gravity? At present, we cannot say.

Another meeting point of general relativity and quantum mechanics is the very beginning of the universe, the big bang itself. The Planck time marks the limit of our ability to speak at all about the evolution of the cosmos. Yet the Planck epoch may be crucial to our understanding of some of our most fundamental questions about the universe, such as how perturbations arose and whether the initial conditions are restricted in some way, or indeed whether they matter at all.

A theory of quantum gravity is required to understand singularities

¹The positron is the anti-electron.

The absence of a quantized theory of gravity is not due to lack of effort by theorists. Many proposals have been put forward, and some progress toward this goal has been made. There as yet exists no complete theory, although the various suggestions may hold pieces of the answer. But before we provide some flavor of the attempts at the unification of gravity and quantum mechanics, we must first discuss a few of the basic ideas of standard quantum theory.

Particles and waves

Although it is a quite modern theory, quantum mechanics has its origins in an old question: is light a particle, or is it a wave? Newton was an early advocate of the corpuscular theory of light, although he recognized that the data were insufficient to decide the issue. However, the observation of such wave phenomena as interference and diffraction, as well as the development of the electromagnetic theory of light in the 19th century, seemed to answer the question most convincingly in favor of the wave. And there things might have remained, were it not for the problem of explaining the blackbody spectrum. Max Planck found he was able to accomplish this by hypothesizing that blackbodies can emit light only in discrete amounts, or *quanta*, with energies proportional to the frequency of the light. The constant of proportionality, symbolized by the letter h , is now known as *Planck's constant*. In 1905 Einstein applied this idea to the *photoelectric effect*, in which a light beam shining upon a metal plate causes an electrical current to flow. The explanation for this effect, which won Einstein his Nobel Prize,² is that light has a particle alter ego, the photon. When a photon of sufficient energy strikes the metal surface, it ejects an electron; these liberated electrons constitute the observed electrical current. The photoelectric effect cannot be explained if light is considered to be a wave; it could only be understood if light took the form of discrete photons. On the other hand, refraction and diffraction cannot be explained if light behaves as a particle; for these phenomena, light must be a wave. Although the same entity may sometimes exhibit wavelike properties, and at other times seem to be a particle, only one such manifestation can be observed at a time.

In quantum mechanics, the blending of particles and waves extends to everything. Not only does light behave both as a particle and a wave, but also electrons, protons, atoms, molecules, and, by extension, even macroscopic objects have both a particle and a wavelike nature. This insight is due to Louis DeBroglie, who proposed in his doctoral thesis in 1924 that a particle could be described by a wave whose wavelength was determined by its momentum. The duality of particle and wave is one of the most counterintuitive ideas of quantum mechanics. Surely, it might seem, an entity should be either a wave or a particle, but not both. However, according to quantum mechanics, *no* experiment can

Particle-wave duality is the foundation of quantum mechanics

²Relativity was thought at the time to be too exotic for the Nobel Prize, which specifies service to humanity.

be devised in which both wave and particle behavior simultaneously appear. This has proven to be true in all experimental tests so far. Under most circumstances, what we envision as an elementary particle, such as a proton or an electron, will act like a particle; but under some conditions, an electron or proton will behave as a wave. It is more unusual to see the wave behavior of a particle such as the electron than it is for light, because the wavelength of light is not small compared to reasonably sized objects, whereas the wavelength of the electron is very short. As a specific example, green light has a wavelength of about 500 nanometers, whereas the wavelength of an electron is only of the order of 0.2 nanometers, about the size of an atom. Light will therefore exhibit wave behavior under many everyday conditions, such as when interacting with air molecules, while an electron will show its wave nature only under more unusual circumstances.

The small wavelength of the electron is exploited by the electron microscope. The resolving power of a microscope, that is, its ability to distinguish two close points, is inversely related to the wavelength of the probe. The shorter the wavelength, the greater the resolving power. Electrons not only have a very short wavelength, but since they are charged and respond to electromagnetic forces, an electron beam can be focused by magnets, just as a beam of visible light can be focused by lenses. In this regard, electrons have a distinct advantage over X-rays, the electromagnetic radiation of comparable wavelength, because X-rays cannot be focused by conventional lenses. In the electron microscope, a beam of electrons is accelerated through an evacuated tube toward the specimen on the stage. Electrons striking the specimen scatter from it, creating an interference pattern that can be refocused at the objective into an image. Electron microscopes are available in several designs and are widely used in research, as well as in the manufacturing of certain items such as semiconductor devices. Electron waves are sufficiently real that an industry has been built up around them.

Quantum mechanics is the physical theory that accommodates this particle-wave duality. Nonrelativistic quantum mechanics is based upon the **Schrödinger equation**, an equation that defines the behavior of an entity called the **wavefunction**. The interpretation of the wavefunction is still not fully unambiguous. The wavefunction must not be regarded as *the “wave of the particle.”* Instead, a very successful and useful interpretation, used every day by physicists working in many subfields, is that the wavefunction describes the *probability distribution* of properties of the system to which it corresponds. A set of attributes, such as energy, momentum, position, and so forth, make up the **quantum state** of a particle; the wavefunction specifies the probability of the particle’s being in a certain state.

The fact that only probabilities, and not absolute certainties, can be assigned to states is ultimately a consequence of the wave nature of particles; this lies at the heart of Heisenberg’s famous uncertainty principle. For example, a wave fills space and therefore its position cannot be unambiguously determined; hence the location of the corresponding par-

The Schrödinger equation describes the evolution of the wavefunction

ticle is uncertain. We can, however, compute the most likely positions for the particle at a given time, by means of the Schrödinger equation. If we measure the position of the particle to greater and greater precision, we find that we can say less and less about its momentum, because for a wave, it is impossible to know both those quantities to arbitrary precision at the same time. To squeeze a wave into a perfectly located position, we cannot use a monochromatic, or single-frequency wave; a perfectly monochromatic wave fills all space. In order to localize a wave, we must add together, or *superpose*, many such monochromatic waves. As we add more and more frequencies, however, we find that the momentum of the wave, which is related to the frequency, is less and less determined. Perfect localization requires an infinite superposition of frequencies, and the momentum becomes completely undefined.

Quantum mechanics is by its very nature a statistical theory. The probabilities can be computed from the wavefunction that a measurement of some variable, such as momentum, position, energy, spin, and so forth, will yield a certain value, but the behavior of a particle is fundamentally unknowable. The limit of our knowledge is defined by Planck's constant, which in MKS units has the value 6.6×10^{-34} joule-seconds. It is the small size of this number which means that we do not see quantum effects in our everyday lives. Yet how do we leap from such a strange, probabilistic realm to the deterministic classical world? If a measurement is performed on an ensemble, that is, on a large number of identical systems, all possible values will be obtained, but the most probable result, called the *expectation value*, will represent the average behavior. For example, it is impossible to predict how long any single atom of uranium will exist before decaying; but the expectation value of the lifetime, measured over a large sample of identical uranium atoms, determines the half-life of the isotope. As the size of the ensemble grows, the expectation value begins to behave more and more like a classical variable; however, no clear-cut demarcation exists at which the quantum world crosses over to the classical.

The wavefunction itself cannot be observed experimentally; only the probabilities computed from it can be measured. It is unclear whether the wavefunction has any physical reality of its own. In the standard interpretation of quantum mechanics, the wavefunction serves merely to define a probability distribution, and only this probability distribution is connected to reality. It is possible to formulate a consistent theory of quantum mechanics, such as that of David Bohm, which differs from standard theory mainly in that the wavefunction does have an objective existence. However, under all conditions achievable on Earth, the predictions of Bohm's theory are identical to those of the standard theory.³ We shall not discuss this interesting digression further, but mention it

The Heisenberg uncertainty principle

The meaning of quantum mechanics

³ Any alternative version of quantum mechanics must reproduce the considerable experimental success of the standard theory in order to be acceptable, in accordance with our usual rules for scientific theories.

only to show that the interpretation of quantum mechanics is still undecided.

The standard view of the wavefunction is called the **Copenhagen interpretation**, because it was formulated by the Danish physicist Niels Bohr. According to Bohr, the wavefunction is merely a mathematical formality that characterizes our state of knowledge about a system; the wavefunction tells us everything that it is possible to know about the particle. In the Copenhagen interpretation, there is a demarcation between the *system* and the *observer*.⁴ Prior to an observation, the wavefunction evolves according to the Schrödinger equation. The variable to be observed, such as momentum, spin, or energy, is described by the probabilities that it will take various values. Each possible value corresponds to a state of the wavefunction; until a measurement is performed, that variable literally has no value, but the wavefunction represents a **superposition of states**, the combination of all possible outcomes for a measurement of that variable. Only when an interaction occurs that demands a particular value for some quantity, such as happens when a measurement is performed, does the observed variable take on a specific value, that which was measured. This rather odd phenomenon is called the **collapse of the wavefunction**. The act of observing caused the wavefunction to assume a state that was previously only a potentiality.

The *two-slit experiment* demonstrates these concepts. In this experiment, a metal plate with two slits, each comparable in thickness to the wavelength of light, is set up perpendicular to the path of a light beam. A screen on the other side of the plate registers the arrival of the light. What is recorded is not two small points of light corresponding to the images of the slits; instead, the light waves emerging from each slit interfere with one other, reinforcing in some places and cancelling in others, to produce an *interference pattern*, bands of alternating light and dark along the screen. The wave behavior here is very similar to the interference patterns resulting when the waves from two sources overlap on the surface of an otherwise still lake.

What if we replace the light source with a source of electrons? This time, we arrange for a single electron to be emitted behind the metal plate and aimed toward it by the attraction of a charged object on the other side. Beyond the plate, we place some kind of detector which we have covered with a phosphorescent material, such as that which coats the display screen of a cathode-ray television set or computer monitor. The impinging of an electron is recorded by a bright glow where it strikes the phosphor. By “observing” the electron, the phosphorescent screen collapses the wavefunction of the electron.

If we repeat this experiment many times, we will find a distribution of positions for the electrons, each occurring with a certain frequency, and all of which can be computed from the wavefunction. After a large

In the Copenhagen interpretation, a measurement causes the collapse of the wavefunction

⁴The term “observer” has the regrettable property of suggesting a conscious entity doing the observing. This is not the case; an “observer” is anything that interacts with the system.

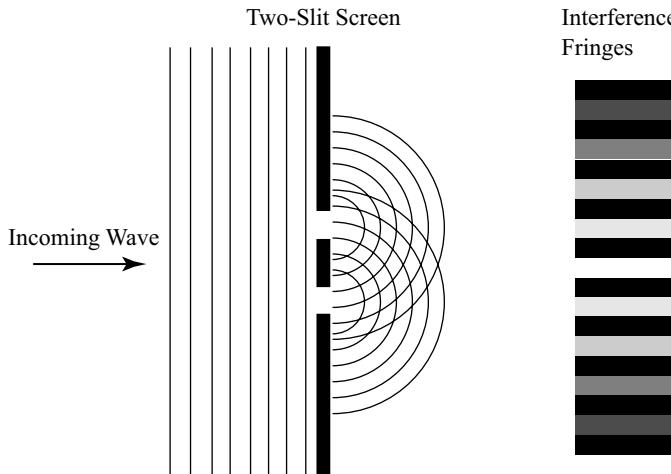


Fig. 17.1 Interference fringes produced by a wave passing through a two-slit screen and striking a detector. Each slit acts as a source for waves. Where the waves from the two slits reenforce, they produce a bright fringe; mutual cancellation produces dark fringes.

number of such measurements of individual electrons, we will find that the locations and repetitions of their measured positions will eventually build into a continuous pattern of interference fringes. The fringes are found to be exactly such as would be produced by a wave that passes through both slits and recombines on the other side. This result is obtained regardless of whether the electrons pass through the slits in a continuous beam or are emitted as individual electrons separated by arbitrarily long intervals of time. More than that, if we placed an appropriate wave detector beyond the screen, we would find an interference pattern no matter how few electrons are admitted into the apparatus; even a *single* electron produces interference fringes! Such fringes could be produced by one solitary electron only if it passed through *both* slits. That is, a lone electron can pass through both slits, interfere with itself, and produce a pattern of fringes! Even if this may seem too bizarre to be possible, it has been confirmed experimentally. In this configuration, we observe the wave aspect of the electron. When electrons pass through both slits, an interference pattern is obtained; the electrons create a wave phenomenon. Something fundamental about the electron's nature is wavelike.

Suppose an experimenter contrived to find out through which slit this single electron really passed. At each slit a detector is placed that registers the passage of every electron, but allows it to proceed through the slit. However, when this experiment is performed the result will be not interference fringes at the screen, but two distinct regions of electron impact points, one for each slit. The act of detecting the electron at one or the other slit collapses the wavefunction. With the electrons localized to one slit or the other, the wave interference that produces fringes is lost.⁵

The two-slit experiment reveals the wave nature of the electron

⁵This phenomenon occurs whether the experimenter personally reads out the slit detector data or not. Again, “measurement” or “observation” does not imply human involvement.

The picture of an electron passing through two separate slits simultaneously is so counter to common sense that it can be quite disturbing. It may be tempting to grant to the invisible world of electrons and photons some strange, almost eerie, properties, but to fall back on the comforting assurance that the real world behaves more logically. In principle, however, quantum mechanics describes the behavior of *all* matter, just as special relativity is the correct theory of dynamics at all speeds. It is simply the case that the wavelength corresponding to any macroscopic object is so tiny that quantum effects are unmeasurable. The quantum wavelength of a thrown baseball, for example, is on the order of 10^{-35} meter.

The tale of Schrödinger's cat

Acceptance of the governance of quantum mechanics over the macroscopic world implies that the quandaries raised by quantum theory carry over to that world as well. If a system unobserved is indeterminate, how does the observer's act of measurement introduce determinacy? We tend to think of the system as microscopic, and the observer as macroscopic. The prejudice is that large objects, such as experimental apparatus, computers, scientists, and the like, must be clearly distinguishable from the bizarre world of barely imaginable particles. But what if the system and the observer are of comparable scale? If quantum mechanics applies to all systems, then quantum effects must in some way control even familiar classical objects; yet quantum behavior seems impossible for everyday entities.

The best-known illustration of this paradox has become part of the folklore of physics, the tale of *Schrödinger's cat*. Suppose a cat were placed into a closed box that is completely isolated from its surroundings. This box contains an elaborate and diabolical contraption. If an atom of some radioactive element decays, the emitted alpha particle trips a Geiger counter, which, by a prearranged switching mechanism, causes a vial to be broken and a poisonous gas to be released into the box. The fate of the cat is tied to a probabilistic quantum effect, specifically, the decay of an atom. Within the box, there is no paradox; the Geiger counter trips or not depending upon whether the decay is observed, and the cat dies or lives accordingly. But consider the system from the point of view of someone outside. The entire setup of box, atom, Geiger counter, and cat is, in principle, a quantum system described by some complicated wavefunction. In any time interval, all that is known is that there is some probability that the atom will decay; so while the cat is unobserved, it is unknown whether it is alive or dead. Adhering strictly to the Copenhagen interpretation forces us to conclude that no measurement has been made while the cat is in the box, and therefore the cat is neither alive nor dead, or else is both alive and dead; that is, the cat–box system represents a superposition of the states “alive” and “dead.” When the outside observer opens the box to observe the state

Schrödinger's cat is a famous thought experiment of quantum mechanics

of the cat, it becomes alive or dead at that moment, according to the probability that the Geiger counter had been tripped.

This is all completely in accord with the laws of quantum mechanics, yet the conclusion seems nonsensical. What might explain the apparent discrepancy between quantum mechanics and well-established common-sense notions that a living being must be either alive or dead, but not both? One suggestion is that a cat in a box is a macroscopic object that is composed of a very large number of microscopic quantum objects, that is, its atoms and molecules, which collude to create the classical behavior we observe. The collective quantum state of such a system would be extraordinarily complex. Writing the “wavefunction of the cat–box system” would be an impossible undertaking, at present. But is it impossible in principle? Perhaps Schrödinger’s equation simply does not apply to such an assemblage, an attitude adopted by many pioneers of the theory, including Schrödinger himself. But this line of argument begs the question, for no clear delimitation has yet been found for the point at which Schrödinger’s equation breaks down in the transition from the microscopic to the macroscopic world.

No literal quantum experiments have ever been carried out with cats, of course. However, many experiments have been performed with great precision upon microscopic entities; the predictions of quantum mechanics and the Copenhagen interpretation are invariably borne out. It is difficult to reconcile Schrödinger’s cat with the classical picture of a cat which is, at any moment, either alive or dead, with a certain probability of its death occurring at any time. It might seem quite reasonable to presume that the states “alive” and “dead” are not quantum states, and thus are not subject to superposition. Yet if quantum mechanics ultimately underlies our macroscopic reality, it must have some validity for apparently classical objects. Perhaps the wavefunction of the cat is so complicated that it is impossible to observe any quantum superposition of states. Or it may be that we cannot prepare a box that is truly so isolated from the rest of the universe that the quantum state describing the total system (cat, radioactive atom, Geiger counter, and poison) can evolve undisturbed by the outside world. Any infringement upon the box by the state of the supposedly external observer might constitute a measurement, which would collapse the wavefunction of the cat. However, none of these alternatives seems to resolve the fundamental paradoxes inherent in the so-called **measurement problem**. What happens in a measurement to cause such a drastic change in the evolution of the system? The collapse of the wavefunction is not described by the Schrödinger equation, but is overlaid upon it as part of the Copenhagen interpretation. When an observation of some variable occurs, the system abruptly ceases to obey the Schrödinger equation; all subsequent measurements of that quantity will continue to yield the same result as long as the system is not otherwise altered. The act of measurement seems to impose reality upon a previously unknowable state; but if that is so, what is reality?

“Schrödinger’s cat” experiments on microscopic entities support the Copenhagen interpretation

The measurement problem is an unresolved issue in quantum mechanics

The collapse of the wavefunction is one of the most vexatious problems of quantum mechanics. Philosophically, most physicists agree that it is at best uncomfortable. Einstein hated it, and his attitude influenced his own efforts to find a unified theory of quantum mechanics and gravity. Much effort is still devoted to analyzing the philosophical underpinnings of quantum mechanics. For operational purposes, however, most physicists set aside aesthetic worries, and use the formal theory of quantum mechanics to make detailed calculations. Some of the predictions of quantum mechanics, including *quantum electrodynamics*, its generalization to include electromagnetics, have been verified to an astounding precision by experiment. We cannot object too much to quantum mechanics on philosophical grounds, then, as it unquestionably describes something very fundamental and deep about the workings of the universe. Perhaps eventually a better interpretation will be found that will clarify these issues.

But there is a more serious objection to the collapse of the wavefunction if we seek to apply quantum mechanics to cosmology. The universe is, by definition, everything observable. The observer is part of the universe. In the Copenhagen interpretation, the collapse of the wavefunction depends upon a clear separation between the observer and the system observed, a distinction which is, obviously, untenable in cosmology. We shall have to put aside this concern for now, however, as it is yet unresolved, and forge ahead.

Quantum cosmology

A direct approach to quantum gravity is to attempt to make a generalization from the Schrödinger equation in one great leap, and to write an equation for the universe as a whole. This is the method that is usually called **quantum cosmology**. How can quantum mechanics, which treats microscopic particles, be extended to the universe as a whole? In quantum mechanics, the Schrödinger equation describes the space and time behavior of the wavefunction for a system, such as a particle, that has some energy, both kinetic and potential. Can this idea be extended to cosmology? The Friedmann equation for \dot{R}^2 (11.17) plays a role much like that of an equation for the energy of a system. This equation can be transformed into a quantum mechanical equation for the evolution of the scale factor. There is no unique way in which to convert the Friedmann equation into a quantum equation; however, one of the best known and most widely applied quantum cosmological equations is the *Wheeler–DeWitt equation*, so named after its developers John Wheeler and Bryce DeWitt.

We shall not write down this equation here, as it is far beyond our scope, but we shall describe some of its consequences. The Wheeler–DeWitt equation requires that the four-dimensional space-time of general relativity be broken into a three-dimensional, purely spacelike, surface and a timelike curve. These, of course, may be identified with

space and time respectively. The decomposition is not unique for a given space-time, however, but depends upon the choice of coordinates, as required by general relativity. The Wheeler–DeWitt equation then describes the evolution of a quantity that bears the rather grandiose title of the **wavefunction of the universe**. Remarkably enough, this quantity is a function of the three-dimensional geometry of the universe and of the matter-energy field; it contains no explicit dependence upon time. It is still unclear whether time even plays a role in the approach to quantum cosmology! Certainly the wavefunction of the universe can change, but what we think of as time may not be the quantity by which we should measure that change.

It is also unclear whether the split into a spacelike surface and a timelike curve is the correct approach to consider. Other possibilities exist and have been studied, some of which may be more promising. Quantum cosmology as a program may be a good approach, but the correct formulation may be lacking. Nevertheless, some interesting results can already be obtained. One of the easier cases to examine is that of a universe which, classically, corresponds to a variant of the de Sitter solution. This model is spherical, that is, closed, and contains a cosmological constant (or, equivalently, a constant vacuum energy density) but no other matter. It is most convenient to study the quantum mechanics of a closed universe; open universes create troublesome extra terms in some of the integrals required, and it is not clear how to handle some of these terms. Moreover, the spherical geometry turns out to be particularly appropriate for quantum cosmology not only because it is highly symmetric and finite, but also because the resulting equations resemble those for a situation familiar to most physicists: that of a particle moving in a potential with a deep minimum, or “well,” in ordinary quantum mechanics.

It happens that the solution for the scale factor of the spherical de Sitter model is a function that has both an exponentially increasing and an exponentially decreasing part. We have already examined the original de Sitter flat-space model, which selected only the exponentially increasing solution. The corresponding spherical geometry, however, admits only a solution with both an expanding and a contracting part. Such a function stretches from $t = -\infty$, at which point $R = \infty$; it contracts as time approaches zero, passes through its minimum size, as specified by the cosmological constant, at $t = 0$, and then expands forever. The minimum size, however, is not zero; this is not a big bang solution, and therefore it contradicts our observations. But the classical solution is not the whole story.

In quantum mechanics, the wavefunction is not confined strictly to the region allowed to it classically. The wavefunction spills over into forbidden regions, though its amplitude becomes small in such regions. Quantum mechanically, then, there is a nonzero probability that a particle may be found in a region where it could not be located classically; this phenomenon is called *quantum tunneling*, and it is exploited in some scientific instruments in use today. In quantum cosmology, the “particle” is the entire universe. Therefore, even though this spherical de Sitter

The Wheeler–DeWitt equation describes the wavefunction of the universe

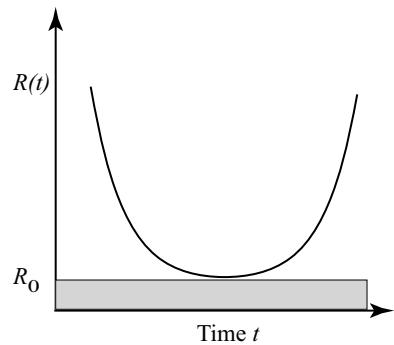


Fig. 17.2 The spherical de Sitter solution. The gray area is the classically forbidden region. Quantum mechanics permits the universe to tunnel from $R = 0$, resuming classical behavior at $R = R_0$.

The universe could have tunneled from $R = 0$

solution cannot classically start from $R = 0$, it can do so quantum mechanically; this universe can tunnel from the state $R = 0$. Beyond the classical minimum value allowed for R , the universe ceases behaving in a quantum-like manner and can continue to evolve classically. This little marvel is a simple consequence of solving the Wheeler–DeWitt equation; the quarrel comes in what to do next.

The Wheeler–DeWitt equation describes only the evolution of the wavefunction; it tells us nothing about the appropriate *boundary conditions*. Different assumptions about these conditions can produce drastically different behaviors. For example, A. Vilenkin has argued that the correct condition should be an outgoing wave, that is, a wave which expands in all directions from $R = 0$, while S. Hawking and J. B. Hartle thought that the appropriate boundary condition should be that there be *no* boundary; the solution should contain both expanding and contracting functions. It is impossible to resolve which, if either, is correct; there simply are no data to help us make such a decision.

Neither does the Wheeler–DeWitt equation address the issue of initial conditions. Different assumptions once again yield different behaviors. It may be that inflation wipes out the initial conditions anyway. It may be that quantum cosmology *creates* the appropriate conditions for inflation to begin. It may be that quantum cosmology sets up initial conditions that would snuff out inflation. It is completely unknown.

What *is* the wavefunction of the universe, anyway? It sounds very pretentious, particularly considering that it arises from a drastic simplification for the universe, a quantum analog of the Friedmann equation. But, despite the necessary oversimplifications, quantum cosmology is beginning to outline important questions that will have to be answered someday in a more complete model. Although the present universe can certainly be said to be evolving in a manner consistent with the purely classical equations, it must nevertheless obey the laws of quantum mechanics. These must become increasingly important as we probe back to the Planck time, and would be important again were the universe to end in a big crunch. But since the universe is apparently unique, how can its wavefunction be given a probabilistic interpretation? And what is the meaning of time within this picture? The answers to these questions remain mysteries.

The nature of time

What is time? This question has troubled philosophers and scientists throughout humanity’s history. Our intuitions say that time is different from space. In space we can travel in any direction, limited only by the capabilities of our modes of transportation. Time, in contrast, seems to be a one-way street, moving inexorably from the past to the future. We remember the past, but can only guess about the future. The past is fixed, unchangeable. The future is indeterminate, mutable, unpredictable. That is at least how we perceive time. But physics takes a

quite different view, one that is not easily reconciled with our experience of time. We have already seen this in our study of relativity. According to special relativity, time and space are joined into space-time. Does special relativity imply that time and space are fully equivalent? Not really; time enters into the metric with a different sign from that of the spatial dimensions, and this is a distinguishing factor, though we have learned that we cannot pin absolute labels onto the passing of time. A particular definition of time depends upon the frame of the observer. Only *proper time*, the time measured by a clock at rest on a given worldline, is invariant. One of the most counterintuitive consequences of the blending of space and time is that simultaneity is not invariant in special relativity, but depends upon the observer: one observer's future may be another observer's past. We insisted upon the preservation of cause and effect, however; an effect could never precede its cause in any frame. The merging of space and time into space-time in special relativity implies that a worldline is not something that creeps forward at some rate, revealing reality as time passes. A worldline is an entity in the space-time. Its future and past are *already there*.

General relativity preserves the basic space-time of special relativity, with the extension in the general theory that the measurement of time, as well as space, depends not only upon velocity but also upon what masses happen to be in the vicinity. An extreme example is the interior of a black hole, where time and space, as defined by the external observer, seem to exchange roles; going forward in time means falling toward the center. But in general relativity as in special, a worldline is a path in the space-time, determined once and for all time by the equations of the theory.

What about the cosmos itself? We have noted previously that the big bang models include a good cosmic time, which is conveniently defined as that proper time kept by an observer at rest with respect to the universe as a whole. But how can we define such an observer during the Planck epoch? Where time and space are themselves subject to uncertainty, we must seek a way of describing space-time events that is free of the arbitrariness of coordinates. It is difficult to know how to begin such a task.

Perhaps time is the wrong marker. Perhaps what we call time is merely a labeling convention, one that happens to correspond to something more fundamental. The scale factor, which is related to the temperature of the universe, could be such a quantity. In our standard solutions, the scale factor, and hence the temperature, is not a steady function of cosmic time. Intervals marked by equal changes in the temperature will correspond to very different intervals of cosmic time. In units of this temperature time, the elapsed interval, that is, the change in temperature, from recombination till the present is less than the elapsed change from the beginning to the end of the lepton epoch.⁶ As an extreme example, if we push temperature time all the way to the big bang, the

What is time?

Temperature may be a better label than time for the evolution of the universe

⁶Figure 12.9 illustrates the concept of “temperature time.”

temperature goes to infinity when cosmic time goes to zero. In temperature units, the big bang is in the infinite past!

In an open universe, the temperature drops to zero at infinite cosmic time, and temperature and cosmic time always travel in *opposite* directions. In a closed universe, on the other hand, there is an infinite temperature time in the future, at some finite cosmic time. A closed universe also has the property, not shared by the open or flat universe, of being finite in both cosmic time and in space. In this case, the beginning and end of the universe are nothing special, just two events in the four-geometry. Some cosmologists have argued for this picture on aesthetic grounds; but as we have seen, such a picture lacks observational support, and has no particular theoretical justification other than its pleasing symmetry.

If we are looking for clues to a physical basis for the flow of time, however, perhaps we are on the right track with temperature. All of the formal theories of physics are time-symmetric. It makes no difference whether time travels backwards or forwards; nothing changes in the equations if we substitute $-t$ for t . Even quantum mechanics makes no distinction between “past” and “future”; the collapse of the wavefunction complicates the picture, but whether that is merely an interpretation of the act of measurement or represents a genuine time asymmetry is currently a matter of debate. The Schrödinger equation itself is as time-symmetric as Newton’s laws. In all of physics, there is only one fundamental law that has a definite time preference: the second law of thermodynamics, which states that entropy increases with time. Newton’s laws, Einstein’s laws, quantum mechanics—all are invariant under a time reversal. Only the second law of thermodynamics proclaims that any process has a direction. How does this fit into the rest of physics?

The second law of thermodynamics is an empirical statement, based originally on observations of steam engines in the 19th century. It seems as though thermodynamics could exist without it. In its modern form, the second law declares that *entropy never decreases, but either remains the same, or increases*. We have stated previously that **entropy** is a measure of the *disorder* of a system. This is certainly true, but more precisely, entropy is related to the total number of macroscopically indistinguishable states that a system can occupy. The more states available, the higher the entropy. For example, the entropy of the air in a room is related to the number of ways in which the air molecules, given the count of molecules present and the total energy available to them, can be physically arranged in space, including rotational freedom of the molecules, such that the macroscopic characteristics of the air are identical. It should be obvious that there are far more ways to arrange molecules that will result, on the large scale, in an even distribution throughout the room, than there are arrangements in which the molecules are clumped in one corner. The evenly distributed state has high entropy (many possible equivalent states), whereas the clumped molecules have low entropy (few possible states). This is why entropy is a measure of disorder; there are generally far more disordered states available than

Entropy and the second law of thermodynamics

there are ordered states. According to the second law, entropy at best remains the same and, in general, increases. Air clumped in the corner of a room, perhaps by means of a piston, will find the state of maximum entropy allowed to it as soon as the opportunity presents itself.

The only processes in which entropy remains the same are *reversible* processes, which are idealizations that do not occur in nature on the macroscopic level, although they are useful theoretically for computing the lower bounds for a process. Real processes are *irreversible*; entropy increases. Irreversibility seems to be intimately related to the direction of time. For instance, a glass tumbler is in a highly ordered state. If it falls and shatters, it enters a more disordered state. Broken shards of glass never spontaneously reassemble themselves into a tumbler; the only way to recreate the tumbler is to melt the fragments and start anew.

The second law applies only to *closed* systems, those in which no energy enters or exits. The universe is certainly a closed system; thus in any process the entropy of the universe as a whole increases. However, local exceptions to the rule of increasing entropy can always be found. As in the example of the recreated tumbler, entropy can be decreased locally by the expenditure of energy. A living creature maintains its highly ordered state only at the cost of enormous consumption of food energy. An automobile converts fuel into mechanical work, specifically, the ordered motions of the pistons and wheels, by extracting the chemical energy of the gasoline. The gasoline that was burned is changed in its composition, and disappears forever as various combustion products, all of which are much less capable of conversion into work. Entropy increases. When a human eats, most of the energy in the food is spent to maintain body temperature; only a fraction goes into driving biochemical processes, while the rest is radiated away into the atmosphere as waste heat. Entropy increases. At death the body's ordered molecules break down into simpler, more disorganized constituents. (They would do so spontaneously, without the aid of bacteria, over a long enough time interval; the bacteria speed up the process and use the energy they extract for their own battle against entropy.)

Entropy and energy

It is not energy that makes the world go 'round. Energy is conserved. The chemical energy released by the burning of gasoline or of food is converted into various forms. Some goes into work; driving pistons and turning wheels, or moving muscles. Some is dissipated by friction; the wheels or feet must overcome friction in order to move, generating heat in the process. Some of the energy goes into maintaining a low-entropy state, such as storing memories in a brain. Some is released as waste heat; through the exhaust of the car, or from the skin. But the total amount of energy, in all forms, is conserved. Not so entropy; entropy increases. From these examples we can also see that another definition of entropy is related to the *capacity to do work*, where *work* is defined strictly in physics as the exertion of a force to produce a motion. A higher entropy state has much less capacity to do work than does a lower entropy state. For example, it is easy to see that uniform heat is a

The increase in entropy determines the arrow of time

higher-entropy form of energy than is kinetic energy. A car moving along at high speed contains a large amount of kinetic energy. If that energy is dissipated into the brake pads, it is distributed into random motions in a huge number of separate molecules. Potentially useful kinetic energy is now spread out in essentially useless, random molecular motions.

Our sense of time moving forward is associated with the change from a state of lower entropy to one of higher entropy. This is why it is so easy to distinguish a motion picture running forward from one that is running backward. Broken glass shards fly together and reassemble themselves into a glass that leaps back onto a table. Crumpled automobiles back away from shattered brick walls, reassembling themselves and the wall in a series of highly coordinated movements. We all know that processes in which order spontaneously increases never occur, so a film shown in reverse strikes us as amusing. In our world, energy must be expended for order to increase; in most natural processes, disorder is created. The **arrow of time** is determined by the inexorable increase in entropy, an increase that is seen in all macroscopic occurrences.

However, there is still a mystery to this. Even with the second law of thermodynamics, there is time symmetry; a system that is disordered today is likely to have been disordered yesterday. That is, if the equations of physics are run backwards, a room full of random air molecules does not revert to an ordered, low-entropy state, but remains in a random, disordered, high-entropy state. To have a sense of the arrow of time, we must start from a low-entropy state. What provides a past and a future in the universe is that it *began* in a state of low entropy; this makes the past distinguishable from the future. Thus the arrow of time is not due to the second law itself. It is due to the *initial condition*.

What is the entropy of the universe? One measure of this quantity is the number of photons. There are perhaps 10^{80} baryons in our visible universe, and about 10^9 photons per baryon. This produces a figure of 10^{89} for the entropy of the cosmic background radiation, and would represent almost all the entropy in the universe were it not for gravity and, more specifically, black holes. Black holes are not completely black; they emit Hawking radiation. Since this radiation is blackbody, a temperature can be assigned to it, and hence to the black hole itself.⁷ The association of a temperature with a black hole leads to a full theory of black-hole thermodynamics, from which an entropy can be derived. Recall that entropy is related to the total number of macroscopically indistinguishable states that a system can have. A black hole is characterized entirely by its mass, charge, and spin; one black hole is indistinguishable from another if both have these same three values. The entropy of a black hole, then, is proportional to the number of states that could have created the black hole. This number is proportional to

⁷Black holes of sizes that are likely to exist are extremely cold; the temperature of a solar-mass black hole is only 10^{-7} K, a remarkably low number, since 0 K is absolute zero, the lowest possible temperature. According to the principle called the third law of thermodynamics, absolute zero cannot be attained but can only be approached arbitrarily closely.

The entropy of the universe

its surface area; the larger the hole, the greater its entropy. Because the surface area is proportional to the mass squared, it follows that the entropy per unit mass increases with mass. From such calculations, it turns out that black holes are the most entropy-laden objects in the universe. If the entire estimated mass of the observable universe collapsed to a black hole, the entropy associated with that black hole would be 10^{123} , a number that is beyond any genuine comprehension and that dwarfs the modest, by comparison, value of 10^{89} of the current universe.

One of the distinguishing features of the black hole is its very strong tidal force. The particles created via Hawking radiation obtain their energy from this tidal force; this implies that the huge entropy of a black hole is somehow tied up with the tidal forces. Roger Penrose has extended this idea to *any* space-time geometry and associated an entropy with the tidal force, which can be computed in a straightforward way from the metric of the space-time. When we apply this to the standard models, we find that the geometries of these space-times have zero tidal force at the big bang. This is quite distinct from the singularity of a black hole, which in contrast has infinite tidal force. Thus the singularity at the beginning of the universe is quite different from that which is to be found in a black hole. If the universe were closed, the final collapse would be essentially a black hole and would have enormously high entropy, which would represent a state quite unlike that from which the cosmos emerged. This alone suggests that the so-called *cyclic model*, in which the universe begins anew following a big crunch, is not likely.

Even an open or flat universe will end in a state of relatively high entropy, due to the increase in entropy as stars burn out and galaxies fade away. This is a remarkable arrangement. Why did the universe begin in such a low-entropy state? Since, as we have asserted, systems seek the state in which their entropy is maximized, we can assume that the natural state of the universe is the aforementioned black hole. This leads us to the conclusion that our universe is special to one part in 10^{123} , in that it actually began from zero entropy. This implies a specialness of the initial conditions to an almost incomprehensible degree. No physical theory known at present is able to account for this phenomenon; the resolution may be buried in the Planck epoch.

The physics of the Planck epoch is inextricably tied to quantum gravity; and it was during the Planck epoch that the conditions were set that resulted in this state of extreme low entropy. Penrose has argued on this basis that the theory of quantum gravity must be a time-asymmetric theory. In this viewpoint, quantum gravity necessarily requires that initial singularities, such as the big bang, be smooth, low-entropy singularities. Such an argument suggests that the apparently special initial conditions are actually part of the laws of physics, and that the relentless march of time is a direct consequence of quantum gravity. Of course, this is only a prescription for a theory, not the theory itself. Still, it is fascinating to contemplate that our perception of the arrow of time might be telling us about the nature of the big bang space-time singularity, from which the macroscopic universe was spawned.

The low entropy of the initial state of the universe may account for the arrow of time

Time travel and many universes

Paradoxes are inherent to time travel

Time is special. Not only does it have a *preferred* direction, but that direction seems inviolable. This does not seem to be demanded by the second law of thermodynamics, with its apparently bland, yet profound, statement about entropy; the second law would seem to require only that time tend to track the increase in entropy. Can we travel backwards in time? All our experience denies this. If it were possible, some severe paradoxes can result. The **grandfather paradox** is one of the best-known of these conundrums. What if a perverse time-traveler visited his own grandfather while he was a baby in his crib and killed the infant? Would the time traveler suddenly disappear, since without his grandfather he could not have eventually come into existence? What if a more benign time-traveler went back in time in order to prevent World War II? Before he left, the world contained the horrors of that war and its aftermath. If the time traveler succeeded in his beneficent mission, would the people killed as a result of the war suddenly return to life? This seems nonsensical.

We have seen that in classical general relativity, any worldline of a material particle must be timelike. However, worldlines do not evolve; they are the complete four-dimensional histories of the particle. Each point, or event, on the worldline represents a particular place at a particular time. If we are to find worldlines that allow time travel, that is, worldlines for which the future lies in the past, we must search for a *closed timelike path* in space-time. Such a worldline is timelike at every event, yet still forms a closed loop. The solutions to the Friedmann equations, and classical black hole solutions, do not allow closed timelike curves. But these are by no means the only solutions to Einstein's equations; might some solutions permit these remarkable worldlines? A new class of closed timelike curves associated with wormholes, discovered by Kip Thorne and collaborators, may seem to be a realization of the dreams of science-fiction writers, but such time machines occur only under extremely special conditions and probably could not be traversed by any real particle. A true worldline is infinitesimally thin; any extended particle describes a world *tube* in space-time. The world tubes found by Thorne are certainly too narrow for macroscopic particles, and may not even be traversable by elementary particles. They also require a pre-existing wormhole that is maintained in a very special state.

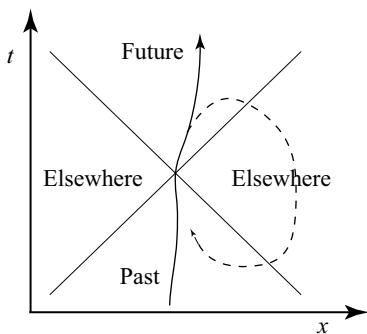


Fig. 17.3 In special relativity a particle's worldline must follow a timelike trajectory. Time travel (dashed line) requires a spacelike trajectory. The curved space of general relativity might permit a closed timelike curve, and this would permit time travel.

One known solution to Einstein's equations that does freely admit closed timelike curves is the *Gödel solution*, found by Kurt Gödel in 1949. The term in Einstein's equations that describes the distribution of stress-energy, which acts as the source of the gravitation, is called, sensibly enough, the *source term*. As it turns out, the Gödel solution corresponds to exactly the same source term as that of Einstein's static cosmology. This vexed Einstein, for it meant that his theory did not perfectly embody Mach's principle; if Mach's principle were strictly upheld, a source distribution would uniquely determine the metric. Gödel's solution proved that this was not the case. In retrospect, it is not surprising;

the Einstein equations are *nonlinear*, and it is a well known property of nonlinear equations that more than one solution may be obtained for the same source.⁸ The Gödel solution is obtained for the same source as the Einstein static solution but with different boundary conditions. In the Gödel solution, it is possible to travel into one's own past, which by itself is a remarkable result. The solution has other curious properties; for instance, it rotates. What does it mean for a universe to "rotate"? With respect to what? To understand this, we must consider the motion of a test particle. Suppose we send a rocket on a track toward a distant galaxy along an inertial worldline. For simplicity, we shall assume that there are no gravitational influences on the rocket other than the galaxy; it simply coasts toward its target. The rocket follows a geodesic, which would, in flat space, correspond to a straight line; eventually the rocket arrives at the galaxy. In the Gödel universe, however, the rocket's path spirals away from the galaxy at which it was aimed.

The Gödel universe does not seem to have much to do with the universe in which we live. No observations have ever detected an overall rotation of the universe, and certainly closed timelike curves are an uncomfortable property at best. The solution demonstrates, however, that the boundary conditions seem to be the way in which Mach's principle is incorporated into general relativity. Those boundary conditions, in the real universe, may specify that classical closed timelike curves are not allowed. This is not a certainty, however.

Might quantum mechanics have anything to say about this? It turns out that in quantum mechanics, the uncertainty principle permits particles to travel into the past so long as causality is not violated. It has even been suggested that antiparticles can be interpreted as particles that are traveling backwards in time! Such an interpretation cannot, however, apply to macroscopic objects. But what if we are wrong, and closed timelike curves do exist? The Robertson–Walker metric is certainly merely an approximation to the universe, valid only on the largest scales; locally, another metric must apply, one which may be greatly complicated and of unknown form. Perhaps such a metric would allow a more varied structure, including closed timelike paths, than the simple metrics we know. If that is true, or else if we find that quantum mechanics somehow permits us to get around the classical limitations, what *would* happen if someone went back into time and killed her grandfather in his cradle?

One means of reconciling this paradox is offered by one of the most exotic conjectures of quantum mechanics, the **many-worlds interpretation**. This interpretation of quantum mechanics, originally proposed by Hugh Everett, was developed to deal with the measurement problem in quantum cosmology. As we have discussed, the Copenhagen interpretation depends upon a distinction between the observer and the system, a distinction that cannot be maintained in quantum cosmology. In the many-worlds interpretation, an infinite number of universes exist. These

The Gödel solution admits closed time-like paths

Can closed timelike curves exist in a realizable space-time?

⁸Only linear equations, in which no variable appears to any power other than the first, can guarantee a unique form of the solution for a fixed source.

Quantum mechanics in the many-worlds interpretation

are not the usual kind of “parallel universes” of science fiction, nor are they the different child universes of the chaotic inflation model. They represent the set of universes in which all possible outcomes of all quantum processes occur. When a measurement is made, no collapse of the wavefunction takes place; rather, the probability of obtaining a given outcome is proportional to the number of universes in which that result is obtained. The issue of the meaning of the collapse of the wavefunction is avoided by requiring all possibilities to occur. As an illustration, return to Schrödinger’s cat. When the box is opened, the universe splits into those in which the cat is alive and those in which it is dead. After the measurement, universes that were once indistinguishable can now be distinguished by whether the cat jumps from the box or not.

The many-worlds interpretation solves not only the measurement problem, but the grandfather paradox as well. If a time-traveler murders his grandfather in the cradle, it merely means that now there are distinguishable universes. In one, the time traveler is never born. This universe, containing the murderer and the dead infant, continues along its own path. In the other universe, the murderer disappears the moment he travels into the past. The objection that the time-traveler and his grandfather are macroscopic, classical objects is inapplicable in this case, because ultimately quantum mechanics must apply to the universe as a whole if we are to solve the measurement problem, and therefore there is no such thing, strictly speaking, as classical behavior.

But if “I” am present in multiple, perhaps infinite, universes, then what are we to make of consciousness? It seems continuous; we remember a past that appears to occur in some linear way. Are there multiple consciousnesses of the same apparent individual, none of which can communicate with the others and each of which regards itself as a single entity? It may be strange to think of Schrödinger’s cat as neither alive nor dead, but is it any more satisfying to suppose that in some universes the cat is alive, while in others it is dead? On the other hand, we have repeatedly stressed that the universe is not bound by our intuition, nor by our sense of aesthetics. If the many-worlds interpretation, or some variant of it, turns out to be the only way in which to fit gravity into quantum mechanics, then we must adjust our common-sense beliefs accordingly.

Whither physics?

Is there any hope that someday we might be able to understand all of these mysteries? It is difficult to know because, at the moment, we are not even sure we are asking meaningful questions. What answers might lie in the immediate future, as more is learned at the frontiers of physics? We had already concluded that we cannot make sense of the beginning of the universe without a theory of quantum gravity, but now we must realize that it goes deeper than that. Quantum mechanics apparently not only accounts for what goes on during the Planck epoch, but also

fundamentally determines the state of the universe, including the arrow of time. Hawking radiation and the entropy of black holes provide tantalizing hints of the wonders that a complete theory of quantum gravity will reveal.

It appears, then, that the arrow of time, which manifests itself in the physics we know only in the form of the second law of thermodynamics, is probably a consequence of quantum gravity. Might quantum gravity determine other things as well? Quantum mechanics is involved with setting the masses of particles; although no complete unified theories have been worked out, masses of some particles can already be computed from theory. The basic mechanism is that of spontaneous symmetry breaking. When a force decouples from the others, certain particles acquire mass, although this phenomenon is still incompletely understood. The more fundamental of the dual roles of mass in physics is that of gravitational charge, with inertia related via the equivalence principle, or whatever underlying quantum principle determines it. It seems likely, then, that quantum gravity will be found to play a major role in determining the masses of elementary particles.

What of other fundamental constants, such as the gravitational constant or Planck's constant? Some approaches to quantum gravity find these constants from the presence of *hidden dimensions*. It seems that even four-dimensional space-time might not be big enough. Many advanced quantum theories require at least ten spatial dimensions in addition to a temporal dimension. The reasons for introducing further dimensions have to do with consistency of various particle models, and are too technical for us to consider here. All the "extra" dimensions are spatial in nature; time remains as a special dimension. These speculated extra dimensions are certainly unobservable at present because they are today *compactified*; that is, they are curled into sizes of the order of the Planck length, 10^{-35}m .⁹ They influence our four-dimensional space-time in that the correct physics is what holds in the higher-dimensional theory. Our world, and the fundamental constants we find to govern it, are obtained in this view by integrating over the hidden dimensions. During the Planck epoch, however, all spatial dimensions might have made comparable contributions, as all would have had the same scale during that interval.

Theories that require hidden dimensions represent a different approach to particle physics, and its unification with gravity, from what we have considered so far. Quantum cosmology is inherently *geometrical*, following the spirit of Einstein's formulation of general relativity. Gravity follows from geometry. Many distinguished physicists have become convinced, however, that the geometrical viewpoint is not leading toward a unified theory, and may even have hindered our progress because it has blinded us to other possibilities. In the alternative approach, particle physics and the unification of forces is fundamental. Gravity arises

The ultimate theory

Quantum gravity may lurk in hidden dimensions

⁹This situation could arise if only three spatial dimensions underwent an inflationary expansion, all the others being left behind as compactified dimensions.

much like the other forces of nature; the exchange of gravitons accounts for gravity, rather than the curvature of space-time. The question still unanswered by such theories is why gravity nevertheless appears to be so geometrical; this is not necessarily a natural consequence of the theories, although further development of the models should illuminate the issue. Perhaps the apparent geometrical nature of gravity is simply an approximation to its true nature, which has yet to be uncovered.

One of the most promising, or at least most interesting, of the non-geometrical approaches is **string theory**. In the standard picture of physics, the fundamental components of nature are extremely tiny, but finite, point particles. In string theory, these fundamental building blocks are instead strings and loops. Such objects can have additional properties and behaviors than can point particles. These strings are quantum objects and their vibrations and interactions occur in a quantum manner. At the level of the Planck scale of lengths, times, and energies, space-time is not continuous but consists of a foam of oscillating and interacting loops and strings. Ultimately, time may not be the linear, smooth function we perceive it to be, but a shifting froth. String theory is the only known theory that unifies all four forces of nature in a finite and self-consistent way. This does not guarantee that the theory is actually a good description of nature, of course. String theory is complex and difficult to understand even for specialists. Nevertheless, it represents a great achievement and provides hope that a unified theory of particle physics and of the four basic forces is within our reach.

String theory

The ultimate question

Throughout most of our study of the universe, we have carefully avoided broaching the question of how the universe began. We have discovered that we are able to describe the evolution of the universe, with considerable success, down to approximately 0.01 second after the beginning. More than that, we have at least some good ideas, and promising hypotheses, to understand the universe to as little as 10^{-43} s after the big bang. If we try to push to times earlier than the Planck time, however, our confidence evaporates, and our ability to say much, beyond the vaguest expressions of our belief in the ultimate unity of the fundamentals of nature, disintegrates. Unless and until we achieve an understanding of the Planck era, it is hardly more than scientific bravado even to speculate about the origin of the universe. It may well be that this subject is beyond science. Still, we cannot resist. After all, it was not so long ago that the nucleosynthesis era, which we now believe we understand quite clearly, was thought to be unreachable by any scientific models. We can contemplate this ultimate question, keeping firmly in mind that any hypothesizing about the origin of the universe can as yet be no better than educated guesswork, and may prove to be nonsense.

The universe is quantum mechanical at its heart. About this there is no doubt. We still do not understand how best to interpret quantum

mechanics, nor do we understand how classical gravity, which works so well for the largest scales of space and time, fits into this indeterminate universe in which we live. Yet if quantum mechanics ultimately governs the life of the universe, is it possible that it might have been responsible for its beginning as well? Quantum mechanics permits the creation of something from nothingness within the universe we observe today. We have suggested that galaxies might trace their ancestry to quantum fluctuations in a quantum field. Might quantum mechanics explain the very origin of the universe itself? If a galaxy could begin as a fluctuation, why not the cosmos as a whole?

Perhaps the fundamental reality is a foam of strings, or else, as another theory suggests, of quantized units of multidimensional space-time with scales of the order of the Planck length and time. Perhaps a fluctuation occurred in at least one region of this foam, causing three spatial dimensions to expand enormously in a kind of inflation; the result would have been a universe that was dominated by an apparently four-dimensional space-time and which then continued to evolve according to the theories we have discussed. This is an intriguing hypothesis, and it has some support from known physics, but as we still have almost no understanding of the Planck era, it must remain an interesting idea that may be all dressed up, but as yet has no place to go.

Quantum gravity is one of the great frontiers of physics at the beginning of the 21st century. When it is achieved, it will certainly answer many of our questions about the universe—but perhaps not all. The end of physics has been pronounced more than once, whenever it was thought that everything was known that could be known. Science is a process of successive approximations. Philosophers sometimes argue that Truth is ultimately unknowable, that our scientific models can never be more than our best description of our own experiences. This may well be correct, but science has enabled us to develop a coherent, self-consistent view of the universe that, in some limit, must show us at least the shape of Truth. We began our exploration of cosmology with the ancient myths and ideas that placed humanity at the center of an unfathomable cosmos. Much of what we have accomplished in the past four centuries has relentlessly removed humanity from any favored place in the cosmos, and has changed the physical universe from a capricious, mysterious realm to a domain that obeys laws we can, at least in some sense, understand. Although the Earth is not the center of the universe, it is representative—all points are equivalent. Thus the conditions required for our existence, the state of nature that has permitted us to form, evolve, and ask these questions, is a state present throughout the universe. At the end of our inquiry we find that, despite all that we have learned, we return to those same questions the ancients asked: Why are we here? Why is the universe here? We are a little closer to answering these questions; perhaps, as the ancients suspected, the answers are linked.

Chapter Summary

At the beginning of time, at the point in the big bang known as the Planck epoch, the universe was sufficiently dense and energetic that all the fundamental forces, including gravity, were merged into one grand force. Although quantum mechanics is generally associated only with the world of the very small, during the Planck epoch the entire observable universe was tiny. Under such conditions, quantum mechanics and gravity must merge into *quantum gravity*. Unfortunately, we do not yet have a theory of quantum gravity, so we speculate on what such a theory might be like, and what it might tell us.

In quantum mechanics everything has a wavelike nature, even those objects we normally consider particles. By the same token, those things that we usually consider waves, such as light, also have a particle nature. The evolution of quantum systems is governed by the Schrödinger equation. However, the Schrödinger equation specifies only the evolution of the probabilities associated with a system. According to the Copenhagen interpretation of quantum mechanics, a system exists in a superposition of states so long as it remains unobserved. An “observation,” which must be understood to refer to any interaction that requires a variable to assume a specific value, collapses the wavefunction. The collapse of the wavefunction means that the system abruptly ceases to obey Schrödinger’s equation; what was previously probabilistic becomes a known quantity. Interpreting what this means is somewhat difficult, given our usual expectations regarding the nature of reality. The story of “Schrödinger’s cat” illustrates the difficulties with standard quantum mechanics as applied to complex systems such as living beings. These problems are only magnified when we seek to apply quantum mechanics to cosmology. Yet we know that it must apply to the universe as a whole, and everything is ultimately quantum.

Cosmology raises questions about the nature of space and time themselves. What is it that provides the “arrow of time,” the perception that we move into the uncertain future and leave behind the unchangeable past? The laws

of physics are time symmetric, meaning that they work the same whether time runs forward or backward. The one exception is the second law of thermodynamics, which states that entropy must increase with time. This means that a complicated system will tend to evolve toward its most probable state, which is a state of equilibrium and maximum disorder. If the sense of the arrow of time arises from the second law, this implies that the big bang must have started in a highly ordered, low-entropy state; the arrow of time results from the universal evolution from this initial state to the final disordered state, be it the big crunch of a closed cosmos or the heat death of the ever-expanding universe. We are led to ask whether quantum gravity might explain why the initial big bang was in a state of low entropy. Perhaps quantum gravity is a theory that is not time symmetric. The second law of thermodynamics may tell us something about the most profound secrets of the universe.

Although we have no established theory of quantum gravity, some promising starts have been made. One of the most studied is string theory, in which reality at the Planck scale of distance and time is described by the quantum oscillations of strings and loops. String theories require that many more spatial dimensions exist than our familiar three. At least ten spatial dimensions exist in these theories, but only three are of cosmic scale; the rest are compactified into “coils” the size of the Planck distance. Thus the very early universe may have undergone a “proto-inflation” in which three spatial dimensions grew into those that make up the observable universe. String theories are not yet well understood and many details remain to be worked out, but so far they can at least unify gravitation and the other fundamental forces in a natural way. Perhaps someday we shall have a better understanding of the mysteries of the Planck scales, and how they determine the universe at the largest scales. Such a discovery would be at least as momentous as general relativity itself.

Key Term Definitions

quantum gravity A unification of gravity and quantum field theory, not yet achieved.

Schrödinger equation The equation that describes the evolution of a nonrelativistic wavefunction.

wavefunction The quantity that obeys the Schrödinger equation. In the Copenhagen interpretation of quantum mechanics, the wavefunction is a mathematical entity that describes the probabilities that the quantum system will assume any of several possible states upon a measurement.

quantum state A particular configuration of quantum properties, for example, energy, spin, momentum, charge, etc., that define a particular system.

Copenhagen interpretation In quantum mechanics, the interpretation of the wavefunction as a description of the probabilities that the state of the system will take on different values.

superposition of states In quantum mechanics, the description of an unobserved system in terms of the probabilities of all possible states.

collapse of the wavefunction In the Copenhagen interpretation of quantum mechanics, the result of an act of measurement, in which the potentialities inherent in the quantum wavefunction take on a specific value, namely, that which is measured.

measurement problem The name for the enigma of how a measurement changes a quantum system into a definite state from one that evolves according to the probabilistic Schrödinger equation.

quantum cosmology A theory that attempts to describe the evolution of the universe in quantum mechanical terms.

wavefunction of the universe A wavefunction that treats the scale factor as a quantum variable and describes its evolution in quantum, rather than classical general relativistic, terms.

entropy A quantitative measure of the disorder of a system. The greater the disorder, the higher the entropy.

arrow of time The direction, apparently inviolable, of the “flow” of time that distinguishes the past from the future.

grandfather paradox The contradictory idea that a time traveler could kill her grandfather while he is an infant in his crib, thus preventing the traveler’s own birth.

many-worlds interpretation An interpretation of the measurement problem in quantum mechanics which holds that each act of measurement causes the universe to split into noncommunicating, parallel, quantum entities.

string theory A theory in which the fundamental structure is not a pointlike particle but is a quantum string, whose vibrations are associated with what we call particles.

Review Questions

- (17.1) What theory is prominent by its absence in our search for the understanding of very extreme conditions? Why do we need such a theory?
- (17.2) What is meant by the phrase “collapse of the wavefunction”? How is this understood within the Copenhagen interpretation? How is it treated in the many-worlds interpretation? How does Schrödinger’s cat illustrate these concepts?
- (17.3) Why is the spherical ($k = 1$) de Sitter solution not an acceptable model within classical physics? What quantum effect occurs to make the spherical de Sitter model potentially compatible with observations?
- (17.4) What does it mean to say that Einstein’s equations or the Schrödinger equation do not distinguish between time running forward or backward? Which law(s) of physics do(es) make a distinction?
- (17.5) Define and discuss the concept of entropy. What is its apparent importance to the arrow of time?
- (17.6) Give an example of a macroscopic phenomenon that looks the same whether time runs forward or

backward. Give an example of a macroscopic phenomenon that is not time symmetric. What distinguishes the two cases?

- (17.7) Give a brief example of the grandfather paradox of time travel. How would you resolve this paradox? Many science fiction stories have centered around travel into the past. If you are familiar with an example, how did the author deal with the subject? (17.9)
- (17.8) Explain how the many-worlds interpretation of quantum mechanics can accommodate the grandfather paradox. Discuss the measurement problem (17.10)
- and explain how the many-worlds interpretation can account for this as well. Do you think that the many-worlds interpretation offers any testable predictions? Is it falsifiable?
- What is meant by the geometrical nature of general relativity? What is the alternative interpretation of gravity?
- What does string theory aim to achieve? Like many newer theories, string theory requires hidden dimensions; why can't we observe them? If they exist, how do they affect the observable universe?

Scientific Notation

Appendix A

Astronomy, even more than most other sciences, demands the use of very large and very small numbers. A convenient notation is essential for dealing with such numbers; specifically, the standard generally known as *scientific notation*. Scientific notation is used routinely in pocket calculators to express large or small numbers, so it is likely that most readers are already familiar with it. A number expressed in this convention has the form

$$N.F \times 10^b,$$

where N is between zero and nine, F is any fractional part, and b is the *exponent*. The number $N.F$ is called the *mantissa*. The exponent simply indicates how many times its *base*, ten in this case, is to be multiplied by itself. For example, 100 is 10×10 . Ten is multiplied by itself twice; therefore,

$$100 = 10^2.$$

In scientific notation, we can write the number 100 as

$$1 \times 10^2,$$

although in this simple example the leading “1” is usually regarded as optional.

Similarly, we can write

$$10^0 = 1 \quad 10^1 = 10 \quad 10^2 = 100 \quad 10^3 = 1000$$

and so forth.

Let us attempt a more general number. Suppose we wish to write 33,500 in scientific notation. We move the decimal place until we achieve a number between zero and nine, counting the number of places we have shifted it. For this case, we must move the decimal point 4 places to the left. The number of places moved indicates the power of ten by which we must multiply the mantissa. It should be clear that $33,500 = 3.35 \times 10,000$; thus we need only write 10,000 in exponent form to obtain

$$33,500 = 3.35 \times 10^4.$$

Scientific notation is almost essential for numbers that are very large; it would be difficult to write a number such as Avogadro’s number, 6.023×10^{23} , without the help of scientific notation. It would be nearly as difficult to follow a long string of “million million million millions” if we were to attempt to write such a number in words. This number is so

large as to be far beyond those which occur in most human activities, and thus it is not easy to conceptualize. Yet is an important constant of nature, describing the number of atoms in a standard quantity of any chemical substance. Large and small numbers create no difficulties for Nature!

So much for very large numbers; what about the small ones? Negative exponents indicate values less than one:

$$10^{-1} = 0.1 \quad 10^{-2} = 0.01 \quad 10^{-3} = 0.001 \quad 10^{-4} = 0.0001$$

Note that in general $10^{-x} = \frac{1}{10^x}$

Multiplication and division are easy with scientific notation. Since 10^3 means $10 \times 10 \times 10$ and 10^2 means 10×10 , it follows that $10^3 \times 10^2$ is equal to 10^5 , or $10 \times 10 \times 10 \times 10 \times 10$. In general:

$$\text{multiplication: } 10^a \times 10^b = 10^{a+b}$$

$$\text{division: } 10^a \div 10^b = 10^{a-b}$$

$$\text{exponentiation: } (10^a)^b = 10^{a \times b}$$

$$\text{taking roots: } \sqrt[b]{10^a} = 10^{a/b}$$

The appropriate operation must also be performed upon the two mantissas. If the result is not between zero and nine, it may be adjusted, with a corresponding change in the exponent. For example:

$$\text{multiplication: } 5 \times 10^4 \times 7 \times 10^6 = 35 \times 10^{10} = 3.5 \times 10^{11}.$$

$$\begin{aligned} \text{division: } 8 \times 10^{10} \div 4 \times 10^5 &= 2 \times 10^5 \\ &8 \times 10^{10} \div 4 \times 10^{20} = 2 \times 10^{-10}. \end{aligned}$$

$$\text{exponentiation: } (3 \times 10^3)^4 = 3^4 \times 10^{12} = 81 \times 10^{12} = 8.1 \times 10^{13}.$$

$$\text{taking roots: } \sqrt[3]{2.7 \times 10^{10}} = \sqrt[3]{27 \times 10^9} = 3 \times 10^3.$$

Units

Appendix B

The *metric system* is used for nearly all scientific purposes. Either of two combinations may be used: the “cgs” (centimeter, gram, second), or “MKS” (meter, kilogram, second) systems. This means that if a length is given in centimeters, then any mass should be expressed in grams, whereas if the length is specified in meters, the mass should be measured in kilograms. The second of time may be used with either choice. Units that adhere to this convention are *consistent*, and can be combined mathematically. Use of units from both systems results in a *mixed* result; this is to be discouraged, although some figures are cited in such units.

Special prefixes indicate powers of ten by which units might be multiplied:

Number	Prefix	Meaning	Example
10^3	kilo	thousand (10^3)	kilogram, kilometer
10^6	mega	million (10^6)	megaparsec, megaton
10^9	giga	billion (10^9)	gigayear, gigahertz
10^{-3}	milli	thousandth (10^{-3})	millimeter, milliamp(ere)
10^{-6}	micro	millionth (10^{-6})	micrometer, microsecond
10^{-9}	nano	billionth (10^{-9})	nanometer, nanosecond

Because metric units were developed with reference to the Earth (the meter was originally intended to be 1/10,000th of the distance from the North Pole to the Equator), they are much too small when applied to astronomical distances and masses; fortunately, scientific notation permits their use. However, astronomers often make use of ‘natural’ units, such as the lightyear, the parsec, and the solar mass. These units have been defined in the text, and are specified in terms of the metric system in Appendix C. “Natural” units are often much more convenient for cosmological quantities, even though they sometimes result in oddly inconsistent units, such as the $\text{km s}^{-1} \text{ Mpc}^{-1}$ usually quoted for the Hubble constant.

Units of any sort are arbitrary, of course. They simply form a set of standards to which we can refer the measurement of physical quantities. The metric units for the most important physical quantities are given below.

- (1) **Time:** The basic unit of time is the second (s). Larger aggregates are the minute, the hour, the day (86,400 s), and the year (3.17×10^7 s). (Note that the year, in astronomy, is strictly 365 days, with each day consisting of 86,400 seconds.)
- (2) **Length:** The units of length are the *meter* and the *centimeter*. The *kilometer* is often used, but remember that it is not a consistent unit within either the cgs or MKS system. (One kilometer is approximately equal to six-tenths, 0.6, of a mile.) Several special units of length are also used in astronomy:
 - (a) The *astronomical unit*, *AU*: This is defined as the distance between the Earth and the Sun, and is approximately 1.5×10^{13} cm.
 - (b) The *lightyear (ly)*: This is defined to be the distance that light travels in one year. It is approximately 9.5×10^{17} cm. Note that the lightyear is a unit of distance, not of time.
 - (c) The *parsec (pc)*: The distance that produces one arcsecond of parallax over the Earth's orbit; it corresponds to about 3.26 ly. The parsec is widely used as a convenient measure of distances between stars. For intergalactic distances, the megaparsec (Mpc) is most common.
- (3) **Velocity:** Standard units of velocity are meters per second (m/s or $m\ s^{-1}$) or centimeters per second (cm/s or $cm\ s^{-1}$). For astronomers a more natural choice is often kilometers per second; $1\ km\ s^{-1} \approx 2200$ miles per hour. For example, the velocity of light is $300,000\ km\ s^{-1}$.
- (4) **Mass:** The units of mass are the gram (g) and the kilogram (kg = 10^3 g). Astronomy also uses units of Earth mass ($M_{\oplus} = 5.97 \times 10^{24}$ kg) or, more usually, Solar mass ($M_{\odot} = 1.99 \times 10^{30}$ kg). For reference, 1 g is the mass of one cubic centimeter of pure water under standard conditions of temperature and pressure.
- (5) **Density:** Density is the amount of mass present in a chosen unit of volume. For example, the density of water under standard conditions is one gram per cubic centimeter ($1\ g/cm^3$ or $1\ g\ cm^{-3}$).
- (6) **Temperature:** Scientists usually measure temperature in *kelvins* (K), or occasionally in degrees Celsius. The size of the unit of each scale is the same, but the kelvin scale locates its zero point at the thermodynamic standard called absolute zero. At absolute zero, all molecules are as stationary as quantum mechanics allows. Absolute zero corresponds to -273°C or to -459°F . The Celsius scale sets its zero at the freezing point of water (+273 K); both positive and negative values are possible. The Celsius scale came to be used in science because it is based upon the phase changes of water, a common substance easily purified and measured in the laboratory. We avoid the Fahrenheit scale, which has smaller units and a rather arbitrary zero point; its zero was based upon a particular combination of ice, salt, and water that is not so easily reproduced consistently in the laboratory.

- (7) **Angular measure:** As the name implies, these are units for the measurements of angles. There are 360 degrees in a full circle, 60 minutes of arc in one degree, and 60 seconds of arc (one arcsecond) in every minute. For example, the Big Dipper (part of the constellation Ursa Major) occupies about 20 degrees on the sky, the full Moon subtends 30 arcminutes, and the eye can barely distinguish two objects separated by one arcminute. The angular size of a dime seen at a distance of two kilometers is approximately one arcsecond.
- (8) **Force:** The standard units of force are the *dyne* (cgs) and the *newton* (MKS). The dyne is one gram-centimeter per second squared, whereas the newton is one kilogram-meter per second squared. Strictly speaking, when we speak of the weight of an object we are actually talking about a force, namely the force exerted on a scale by a mass accelerated by the gravitational field of the Earth. The kilogram is a unit of mass, *not* of weight; however, the English unit *pound* is a unit of weight, not of mass. In countries that utilize the metric system the kilogram is nearly always used for weight; for some reason, the newton is not widely used in everyday life. Near the surface of the Earth, however, there is little practical distinction between mass and weight.
- (9) **Energy and power:** The standard units of energy are the *erg* (cgs) or the *joule* (MKS). The joule is much larger than the erg, with $1 \text{ J} = 10^7 \text{ erg}$. One joule corresponds to the energy obtained from dropping 1 kg (such as a small bag of sugar) from a height of 10 cm (close to 4 inches). The more familiar unit, the *watt* (W), is actually a unit of **power**, which is the rate of energy production or release per second. Specifically, 1 watt is equal to 1 joule per second. Thus a 100-watt light bulb expends 100 joules of energy every second. The hybrid unit *kilowatt-hour*, which appears on most electric bills in the United States, consists of a unit of power multiplied by a unit of time, and thus is itself a unit of energy; customers are billed for the total energy consumed over some interval of time, not for “power” *per se*. In astronomy, a natural unit of power is the luminosity of the Sun, $1 L_\odot = 3.9 \times 10^{26} \text{ W}$, or $3.9 \times 10^{33} \text{ erg s}^{-1}$.

Units conversions

Many people find units conversions intimidating. One way to prevent mistakes is to remember that the symbols for units behave exactly like algebraic variables. It is possible to write an equation containing units, and to cancel like (and *only* like) units appropriately. For example, suppose it is desired to convert kilometers into miles. We have the fundamental equation that

$$0.61 \text{ mile} = 1 \text{ km.}$$

This is an equality, so we may write

$$\frac{0.61 \text{ mile}}{1 \text{ km}} = 1$$

or its reciprocal

$$\frac{1 \text{ km}}{0.61 \text{ mile}} = 1.$$

Algebraically, we may always multiply by "1" in some form. Suppose we wish to know

$$56 \text{ miles} = ? \text{ km.}$$

If we always keep in mind that we must manipulate the problem into the form such that the units we wish to eliminate will cancel algebraically, we can see that we obtain

$$56 \text{ miles} \times \frac{1 \text{ km}}{0.61 \text{ mile}} = 92 \text{ km.}$$

This method works even when compound units are to be converted. For instance, suppose we must convert a density from g cm^{-3} to kg m^{-3} . As a specific example, the density of water is 1 g cm^{-3} ; what is its density expressed in kg m^{-3} ?

We begin with the equalities

$$1000 \text{ g} = 1 \text{ kg},$$

$$100 \text{ cm} = 1 \text{ m.}$$

From the second of these, we find that

$$100^3 \text{ cm}^3 = 1^3 \text{ m}^3$$

or

$$10^6 \text{ cm}^3 = 1 \text{ m}^3.$$

Thus

$$\frac{1 \text{ g}}{\text{cm}^3} \times \frac{1 \text{ kg}}{10^3 \text{ g}} \times \frac{10^6 \text{ cm}^3}{1 \text{ m}^3} = 1000 \text{ kg m}^{-3}.$$

Hence the density of water is 1000 kg m^{-3} .

Physical and Astronomical Constants

Appendix C

Quantity	Symbol	Value
Speed of light	c	$3.00 \times 10^8 \text{ m s}^{-1}$
Gravitational constant	G	$6.67 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}$
Boltzmann's constant	k	$1.38 \times 10^{-23} \text{ J K}^{-1}$
Planck's constant	h	$6.63 \times 10^{-34} \text{ J s}$
Proton mass	m_p	$1.67 \times 10^{-27} \text{ kg}$
Electron mass	m_e	$9.11 \times 10^{-31} \text{ kg}$
Mass of Earth	M_\oplus	$5.98 \times 10^{24} \text{ kg}$
Radius of Earth	R_\oplus	$6.37 \times 10^6 \text{ m}$
Mass of Sun	M_\odot	$1.99 \times 10^{30} \text{ kg}$
Radius of Sun	R_\odot	$6.96 \times 10^8 \text{ m}$
Astronomical unit	AU	$1.50 \times 10^{11} \text{ m}$
Light year	ly	$9.46 \times 10^{15} \text{ m}$
Parsec	pc	$3.09 \times 10^{16} \text{ m}$

This page intentionally left blank

Glossary

absolute zero: The lowest possible temperature, attained when a system is at its minimum possible energy. The Kelvin temperature scale sets its zero point at absolute zero (-273.15° on the Celsius scale, and -434.07° on the Fahrenheit scale).

absorption spectrum: A spectrum consisting of dark lines superimposed over a continuum spectrum, created when a cooler gas absorbs photons from a hotter continuum source.

acceleration: A change of velocity with respect to time.

accretion disk: A disk of gas that accumulates around a center of gravitational attraction, such as a white dwarf, neutron star, or black hole. As the gas spirals in, it becomes hot and emits light or even X-radiation.

active galactic nucleus (AGN): An unusually bright galactic nucleus whose light is not due to starlight.

active galaxy: A galaxy whose energy output is anomalously high. About 1% of galaxies are active. Most contain an AGN at their cores.

amplitude: See wave amplitude.

angular size: The angle subtended by an object on the sky. For example, the angular size of the full Moon is 30 arcminutes.

anthropic principle: The observation that, since we exist, the conditions of the universe must be such as to permit life to exist.

anthropocentrism: The belief that humans are central to the universe.

anthropomorphism: The projection of human attributes onto nonhuman entities such as animals, the planets, or the universe as a whole.

antimatter: Particles with certain properties opposite to those of matter. Each matter particle has a corresponding *antiparticle*. The antiparticle has exactly the same mass and electric charge as its partner. When a particle combines with its antiparticle both are annihilated and converted into photons. (*See also* baryogenesis.)

arrow of time: The direction, apparently inviolable, of the “flow” of time that distinguishes the past from the future.

Astronomical Unit (AU): The mean distance from the Earth to the Sun.

astronomy: The study of the contents of the universe beyond the Earth.

atom: The smallest component of matter that retains its chemical properties. An atom consists of a nucleus and at least one electron.

atomic number: The number of protons present in the nucleus of an atom. This determines its elemental identity.

baryogenesis: The creation of matter in excess of antimatter in the early universe. Only the relatively few unmatched matter particles survived to make up all subsequent structures.

baryon: A fermionic particle consisting of three quarks. The most important baryons are the proton and the neutron.

baryon number conservation: The principle that the number of baryons must remain the same in any nuclear reaction.

biased galaxy formation: The theory that the distribution of galaxies is not representative of the overall matter distribution in the universe because galaxies form preferentially from anomalously over-dense dark-matter perturbations.

big bang: The state of extremely high (classically, infinite) density and temperature from which the universe began expanding.

big crunch: The state of extremely high density and temperature into which a closed universe will recollapse in the distant future.

Birkhoff's theorem: A theorem of general relativity which states that all spherical gravitational fields, whether from a star or from a black hole, are indistinguishable at large distances. A consequence of this is that purely radial changes in a spherical star do not affect its external gravitational field.

black hole: An object that is maximally gravitationally collapsed, and from which not even light can escape.

black hole thermodynamics: The theory that permits a temperature and an entropy to be defined for black holes.

blackbody: A perfectly absorbing (and perfectly emitting) body.

blackbody radiation: A special case of thermal radiation, emitted by a blackbody and characterized by thermal equilibrium of the photons. A blackbody spectrum is completely determined by the temperature of the emitter.

blueshift: A shift in the frequency of a photon toward higher energy.

boost factor: The quantity $\Gamma = 1/\sqrt{1-v^2/c^2}$ that relates measurements in two inertial frames according to special relativity.

boson: A class of elementary particles whose spin is an integer multiple of a fundamental quantized value. The major function of bosons is to mediate the fundamental forces. The best-known boson is the photon.

bottom-up structure formation: The theory that small structures, perhaps galaxies or even smaller substructures, form first in the universe, followed later by larger structures.

brown dwarf: A substellar object that is near, but below, the minimum mass for nuclear fusion reactions to occur in its core.

carrier boson: A particle that carries one of the fundamental forces between other interacting particles. For example, the carrier boson for the electromagnetic force is the photon.

CBR: See cosmic background radiation.

Cepheid variable: A type of variable star whose period of variation is tightly related to its intrinsic luminosity.

Chandrasekhar limit: The maximum mass, approximately $1.4M_{\odot}$, above which an object cannot support itself by electron degeneracy pressure; hence it is the maximum mass of a white dwarf.

chaotic inflation: A model in which many distinct universes form from different regions of a mother universe, with some inflating and others perhaps not.

charge: The fundamental property of a particle that causes it to participate in the electromagnetic force.

closed universe: A standard model with a spherical three-dimensional spatial geometry. Such a universe is finite in both space and time, and recollapses.

cold dark matter: A form of nonbaryonic dark matter that has low energy and low particle velocities at the time it decouples from other matter early in the history of the universe, and whose mass plays a key role in cosmic structure formation. Cold dark matter leads to bottom-up structure formation.

cold dark matter model: A model of structure formation in which an exotic particle whose energy is low at the time it decouples from other matter is responsible for structure formation.

collapse of the wavefunction: In the Copenhagen interpretation of quantum mechanics, the result of an act of measurement, in which the potentialities inherent in the quantum wavefunction take on a specific value, namely, that which is measured.

collisionless damping: The tendency of weakly interacting (collisionless) matter to smooth out gravitational perturbations by freely streaming from over-dense to underdense regions.

comoving coordinates: Coordinates fixed with respect to the overall Hubble flow of the universe, so that they do not change as the universe expands.

Compton wavelength: The quantum wavelength of a particle with a highly relativistic velocity.

concordance model: A model of the universe that has the best overall agreement with data from a variety

of observations, including redshift–distance tests, cosmic background radiation fluctuations, and big bang nucleosynthesis calculations.

conservation of angular momentum: The principle that the angular momentum of a system (the momentum of rotation about a point) remains the same as long as no external torque acts.

conservation of energy: The principle that the total energy of a closed system never changes, that energy is only converted from one form to another. This principle must be enlarged under special relativity to include mass-energy.

conservation of matter: The principle that matter is neither created nor destroyed. This principle is only approximately true, since it is violated by special relativity.

conservation of momentum: The principle that the linear momentum of a system (in Newtonian mechanics, mass times velocity) remains the same as long as no external force acts.

consistent: Of a scientific theory: containing and extending an earlier well-supported theory, e.g. general relativity is consistent with Newtonian gravity.

coordinate singularity: A location at which a particular coordinate system fails, such as the Schwarzschild metric coordinates at the Schwarzschild radius of a black hole.

coordinates: Quantities that provide references for locations in space and time.

Copenhagen interpretation: In quantum mechanics, the interpretation of the wavefunction as a description of the probabilities that the state of the system will take on different values.

Copernican principle: The principle that the Earth is not the center of the universe.

Copernican revolution: The revolution in thought resulting from the acceptance of the heliocentric model of the solar system.

correlation function: A mathematical expression of the probability that two quantities are related. In cosmology, the correlation function indicates the probability that galaxies, or clusters of galaxies, will be found within a particular distance of one another. The correlation function provides a quantitative measure of the clustering of galaxies (or of clusters).

cosmic background radiation (CBR): The blackbody radiation, now mostly in the microwave band, consisting of relic photons left over from the very hot, early phase of the big bang.

cosmic censorship: The principle that singularities are never “naked,” that is, they do not occur unless surrounded by a shielding event horizon.

cosmic distance ladder: The methods by which increasing distance is measured in the cosmos. Each depends on a more secure technique (or “rung”) used for smaller distances.

cosmic time: A time coordinate that can be defined for all frames in a homogeneous metric, representing the proper time of observers at rest with respect to the Hubble flow. In a big bang model, this coordinate marks the time elapsed since the singularity.

cosmological constant: A constant introduced into Einstein’s field equations of general relativity in order to provide a supplement to gravity. If positive (repulsive), it counteracts gravity, while if negative (attractive), it augments gravity. It can be interpreted physically as an energy density associated with space itself. *See also* dark energy.

cosmological principle: The principle that there is no center to the universe, that is, that the universe is isotropic on the largest scales, from which it follows that it is also homogeneous.

cosmological redshift: A redshift caused by the expansion of space.

cosmology: The study of the origin, evolution, and behavior of the universe as a whole.

critical density: That density that just stops the expansion of space, after infinite cosmic time has elapsed. In the standard models, the critical density requires that the spatial geometry be flat.

crucial experiment: An experiment that has the power to decide between two competing theories.

curvature constant: A constant (k) appearing in the Robertson–Walker metric that determines the curvature of the spatial geometry of the universe.

dark age: The era, lasting hundreds of millions of years, between the epoch of recombination and the onset of star and galaxy formation.

dark energy: The term given to the energy that is responsible for the overall acceleration of the universe. Possible dark energies include a cosmological constant Λ , a nonzero vacuum energy, or otherwise unknown forms of energy dubbed *quintessence*.

dark halo: A massive aggregation of nonluminous matter of unknown kind that surrounds and envelopes galaxies.

dark matter: Matter that is invisible because it emits little or no light. Most generally, dark matter includes both ordinary baryonic matter and any exotic forms of matter. Sometimes used in a more restrictive sense to refer to nonbaryonic gravitating mass, as in hot dark matter or cold dark matter. Most of the mass of the universe is dark.

data : The outcome of a set of measurements from which inferences may be drawn, theories constructed, etc.

de Sitter model: A model of the universe that contains no matter, but only a positive cosmological constant. It expands exponentially forever.

deceleration parameter: A parameter (q) that measures the rate of change with time of the Hubble constant.

density parameter: The ratio of the actual mass-energy density of the universe to the critical density. Also called omega (Ω).

deuterium: An isotope of hydrogen whose nucleus contains one proton and one neutron.

distance ladder: See cosmic distance ladder.

dynamical method: A method of measuring the mass of a galaxy, cluster, or even the universe that makes use of the gravitational interactions of two or more bodies.

Doppler effect: The change in frequency of a wave (light, sound, etc.) due to the relative motion of source and receiver.

Einstein equivalence principle: The principle that *all* physical laws, not just those of mechanics, are the same in all inertial and freely falling frames of reference.

Einstein-de Sitter model: The flat ($k = 0$), pressureless standard model of the universe.

electromagnetic force: The force between charged particles, which accounts for electricity and magnetism. One of the four fundamental forces of nature, it is carried by photons and is responsible for all observed macroscopic forces except gravity.

electromagnetic spectrum: The full range of light wavelengths or frequencies, from low energy radio waves to high energy gamma rays.

electron: An elementary lepton with a negative charge. One of the components of atoms, the electrons determine the chemical properties of an element.

electron degeneracy: A condition of matter in which all quantum states available to the electrons are filled.

electron degeneracy pressure: A form of pressure arising from electron degeneracy; the electrons resist being forced closer together because of the exclusion principle.

electroweak interaction: The unified electromagnetic and weak forces. Also called the electroweak force.

element: A particular type of atom, with specific atomic number and chemical properties. The smallest unit into which matter may be broken by chemical means.

ellipse: A geometric figure generated by keeping the sum of the distance from two fixed points (the foci) constant.

elliptical galaxy: A galaxy whose shape is roughly spheroidal or ellipsoidal. Most ellipticals contain little dust or gas, and show no evidence of recent star formation.

elsewhere: Those events in space-time that cannot be causally connected to a given event.

emission distance: The distance to the source of light at the time the light was emitted.

emission spectrum: A spectrum consisting of bright lines, created when a hot gas emits photons characteristic of the elements of which the gas is composed.

energy: The capacity to perform work, where *work* is defined as the exertion of a force to produce a displacement.

entropy: A quantitative measure of the disorder of a system. The greater the disorder, the higher the entropy.

equal density epoch: That interval in the early history of the universe when the gravitational contributions of matter and radiation were approximately equal.

equilibrium: A balance in the rates of opposing processes, such as emission and absorption of photons, creation and destruction of matter, etc.

equivalence principle: The complete equality of gravitational and inertial mass, gravity and acceleration, and the identification of freefalling frames with inertial frames. (See also Einstein equivalence principle, and Newtonian equivalence principle.)

ergosphere: The region of a rotating Kerr black hole between the static surface and the event horizon.

escape velocity: The minimum velocity required to escape to infinity from the gravitational field of an object.

Euclidean geometry: Flat geometry based upon the geometric axioms of Euclid.

event: A point in four-dimensional space-time; a location in both space and time.

event horizon: A surface that divides space-time into two regions; that which can be observed, and that which cannot. The Schwarzschild radius of a non-rotating black hole is an event horizon.

exclusion principle: The property that fermions of the same type which can interact with each other can not simultaneously occupy the same quantum state.

experiment: A controlled trial for the purpose of collecting data about a specific phenomenon.

explanatory power: The ability of a scientific hypothesis to account for known data.

extinction: In astronomy, the removal of light from a beam by whatever means, such as absorption and scattering.

false vacuum: A metastable state in which a quantum field is zero, but its corresponding potential is not zero.

falsifiable: Of a scientific hypothesis: leading to the possibility of performing an experiment that would disprove, or falsify, the hypothesis.

fermion: A class of elementary particles whose spin is a half-integer multiple of a fundamental quantized value. Fermions make up matter. The best-known fermions are protons, neutrons, electrons, and neutrinos. Fermions obey the exclusion principle.

field: A mathematical representation of a quantity describing its variations in space and/or time.

first acoustic peak: The longest wavelength maximum in the temperature fluctuations of the cosmic background radiation. This wavelength corresponds to a pressure wave in the early universe that has completed half an oscillation cycle, and hence has reached maximum compression, at the time of recombination.

fission: The splitting of a heavy atomic nucleus into two or more lighter nuclei.

flat geometry: Geometry in which the curvature is zero; ordinary Euclidean geometry.

flat universe: A model whose three-dimensional spatial geometry is flat.

flatness problem: The observed fact that the geometry of the universe is very nearly flat, which is a very special condition, without an explanation of why it should be flat.

flux: The amount of some quantity, e.g. energy, crossing a unit area per unit time.

force: That which produces an acceleration.

frame of reference: The coordinate system to which a particular observer refers measurements.

freefall: Unrestrained motion under the influence of a gravitational field.

frequency: *See* wave frequency.

Friedmann equation: The equation that describes the evolution of the cosmological scale factor of the Robertson-Walker metric.

Friedmann model: A cosmological model that is isotropic, homogeneous, and governed by the Friedmann equation.

fundamental forces: The four forces (strong, weak, electromagnetic, and gravitational) that account for all interactions of matter.

fusion: The joining of two or more lighter elements to create a heavier nucleus.

future: Those events that could be influenced by a given event.

galactic cannibalism: The process of galaxy merger in which a large galaxy disrupts and assimilates a smaller galaxy.

galaxy: A large, gravitationally bound system of stars, star clusters, and interstellar matter.

galaxy cluster: A group of galaxies that are mutually gravitationally bound.

Galilean relativity: The transformation from one inertial frame of reference to another in the limit of very small velocities and very weak gravitational fields.

gauge boson: *See* carrier boson.

geocentric: Taking the Earth to be the center, for example of the solar system.

geodesic: In geometry, that path between two points/events which is an extremum in length. In some geometries, such as Euclidean, the geodesics are the shortest paths, whereas in others, such as in the space-time geometries appropriate to general relativity, the geodesics are the longest paths.

globular cluster: An aggregation of approximately 100,000 stars. Halos of globular clusters orbit many galaxies. Some globular clusters are thought to be among the oldest structures in the universe.

gluon: A hypothetical particle that binds quarks together into hadrons.

grand unified theory: A member of a class of theories that seek to explain the unification of the strong, weak, and electromagnetic forces.

grandfather paradox: The contradictory idea that a time traveler could kill her grandfather while he is an infant in his crib, thus preventing the traveler's own birth.

gravitational constant: A fundamental constant of nature, G , which determines the strength of the gravitational interaction.

gravitational lens: A massive object that causes light to bend and focus due to its general-relativistic effect upon the space-time near it.

gravitational radiation: The emission of gravitational waves by the creation of a gravitational field which changes in time. Also: the waves (*see* gravitational wave) so radiated.

gravitational redshift: A shift in the frequency of a photon to lower energy as it climbs out of a gravitational field.

gravitational wave: A propagating ripple of space-time curvature that travels at the speed of light.

graviton: A hypothetical massless boson that is the carrier of the gravitational force.

gravity: The weakest of the four fundamental forces; that force which creates the mutual attraction of masses.

hadron: A class of particles that participate in the strong interaction. Hadrons consist of those particles (baryons, mesons) which are composed of quarks.

hadron epoch: That interval in the early history of the universe after the quarks had condensed into hadrons, and before the temperature dropped below the threshold temperature for protons.

half-life: The interval of time required for half of a sample of a radioactive material to decay.

Harrison-Zel'dovich spectrum: A proposed spectrum for the matter perturbations in the early universe that later became the observed structure. The Harrison-Zel'dovich spectrum is scale-free, i.e. perturbations of all sizes behave in the same way.

Hawking radiation: Emission of particles, mostly photons, near the event horizon of black holes due to the quantum creation of particles from the gravitational energy of the black hole.

heat: A form of energy related to the random motions of the particles (atoms, molecules, etc.) which make up an object.

heat death: The fate of the open universe models in which the temperature drops toward zero, stars die out, black holes evaporate from Hawking radiation, entropy increases, and no further energy is available for any physical processes.

heliocentric: Taking the Sun to be the center, for example of the solar system.

Higgs boson: A hypothetical particle that plays an important role in Grand Unified Theories. The Higgs boson would be associated with processes leading to baryogenesis, and might play a role in endowing all particles with mass.

homogeneity: The property of a geometry that all points are equivalent.

horizon: Any surface that demarcates events which can be seen from those which cannot be seen.

horizon problem: The conflict between the observed high uniformity of the cosmic background radiation and the fact that regions of the sky separated by an angular size of more than approximately one degree could not have been in causal contact at the time of recombination.

hot dark matter: A form of nonbaryonic dark matter that has high energy and high particle velocities at the time it decouples from other matter early in the history of the universe. Such matter tends to clump gravitationally into large galaxy-cluster-sized structures initially, leading to top-down structure formation.

hot dark matter model: A model of structure formation in which a particle whose energy is high at the time it decouples from other matter is responsible for the origin of large-scale structure.

Hubble constant: The constant of proportionality (H) between recession velocity and distance in the Hubble law. It is not actually a constant, because it can change with time over the history of the universe.

Hubble expansion: The separation of galaxies due to the expansion of space, not due to their individual gravitational interactions.

Hubble flow: The separation of galaxies due only to the overall expansion of space.

Hubble law: The relationship between recession velocity and distance, $v = H\ell$, for an isotropic, expanding universe.

Hubble length: The distance traveled by light along a straight geodesic in one Hubble time, $D_H = ct_H$.

Hubble sphere: A sphere, centered about any arbitrary point, whose radius is the Hubble length. The center of the Hubble sphere is not a “center” to the universe, because each point has its own Hubble sphere. The Hubble sphere approximately defines that portion of the universe that is observable from the specified point at a specified time.

Hubble time: The inverse of the Hubble constant, $t_H = 1/H$. The Hubble time, also called the Hubble age or the Hubble period, provides an estimate for the age of the universe.

hydrostatic equilibrium: The balance between gravity and gas pressure in an object such as a star.

hyperbolic geometry: A geometry that has negative constant curvature. Hyperbolic geometries cannot be fully visualized, because a two-dimensional hyperbolic geometry cannot be embedded in a three-dimensional Euclidean space. However, the lowest point of a saddle, that point at which curvature goes both “uphill” and “downhill,” provides a local representation.

hypothesis: A proposed explanation for an observed phenomenon. In science, a valid hypothesis must be based upon data and must be subject to testing.

ideal gas: A gas in which the mutual interactions of the gas particles are negligible, except for their momentary collisions. The pressure is determined by the ideal gas law.

ideal gas law: The formula that relates temperature, pressure, and volume for an ideal gas. Nearly all real gases obey the ideal gas law to very high temperatures and pressures, even those found in the interiors of stars.

inertia: That property of an object which resists changes in its state of motion.

inertial force: A force arising from the acceleration of an observer’s frame of reference.

inertial motion: Motion free of any force, that is, constant velocity motion.

inertial observer: An observer occupying an inertial frame of reference.

inertial reference frame: A reference frame in which a free particle experiences no force.

inflation: A period of exponential increase in the cosmic scale factor due to a nonzero vacuum energy density, which occurs early in the history of the universe in certain cosmological models.

inflaton: The generic name of the unidentified particle that may be responsible for an episode of inflation in the very early universe.

initial mass function: The theoretical function describing the number of stars for each given mass that will be produced in an episode of star formation.

interference: The interaction of two waves in which their amplitudes are reenforced and/or cancelled.

interference fringes: A pattern of alternating reinforcement and destruction caused by the interference of two or more waves.

interferometer: A device that carries out some measurement by detecting wave interference.

interstellar medium: Gas, dust, bits of ice, etc. that fill the space between the stars. Nearly all of the interstellar medium is hydrogen and helium gas, with hydrogen most abundant.

invariance: The property of remaining unchanged under a transformation of the frame of reference or the coordinate system.

ion: An atom that has gained or lost an electron and thereby acquired an electric charge. (Charged molecules are usually called radicals, not ions.)

irregular cluster: A cluster of galaxies with no particular shape. Irregular clusters often contain many spiral galaxies.

irregular galaxy: A galaxy with an ill-defined, irregular shape. Many irregulars are interacting or even colliding with other galaxies, which may account for their disorganized appearance.

isotope: One of the forms in which an element occurs. One isotope differs from another by having a different number of neutrons in its nucleus. The number of protons determines the elemental identity of an atom, but the total number of nucleons affects properties such as radioactivity or stability, the types of nuclear reactions, if any, in which the isotope will participate, and so forth.

isotropy: The property of sameness in all directions, as in an isotropic geometry.

Kepler's Laws: The three laws of planetary motion discovered by Johannes Kepler.

Kerr metric: The metric that describes the space-time around a rotating black hole.

kinematical method: A method of measuring the mass density of the universe indirectly, by means of overall parameters of the universe such as its expansion rate. Kinematic methods exploit the fact that expansion rate, deceleration parameter, density, and curvature are not completely independent quantities, but are related by the Friedmann equations, possibly extended to include a cosmological constant.

kinetic energy: The energy associated with macroscopic motion. In Newtonian mechanics, the kinetic energy is equal to $\frac{1}{2}mv^2$.

lambda density parameter: Analogous to the matter density parameter, this term, written Ω_Λ , measures the relative importance of the Λ term compared to the critical value that would correspond to a flat universe.

law: In scientific usage, a theory that has become particularly well confirmed and well established.

law of inertia: Another name for Newton's first law of motion.

Lemaître model: The cosmological model developed by Georges Lemaître, which contains a positive cosmological constant, uniform matter density, and spherical spatial geometry.

length contraction: An apparent contraction of the length of an object in motion relative to a given observer, caused by the Lorentz transformation from one frame to another.

lepton: A member of a class of fermionic particles that do not participate in the strong interaction. The best-known lepton is the electron.

lepton epoch: The interval in the early history of the universe when leptons dominated.

lightcone: The surface representing all possible paths of light that could arrive at or depart from a particular event.

lightlike: Of a space-time interval: capable of being traversed only by a massless particle such as a photon. A lightlike, or null, space-time interval is zero. *See also* null.

lightyear (ly): A measure of distance equal to that traveled by light in one year.

line radiation: Radiation of a particular wavelength produced by an electron moving from one orbital to another of lower energy. (*See also* emission spectrum.)

Local Group: The small cluster of galaxies of which our Galaxy and the Andromeda Galaxy are prominent members.

long range force: A force that does not become equal to zero within any finite distance. The long-range forces are gravity and electromagnetism, both of which decrease as R^{-2} with increasing distance R .

lookback time: The time required for light to travel from an emitting object to the receiver.

Lorentz contraction: *See* length contraction.

Lorentz transformation: The transformation, valid for all relative velocities, which describes how to relate coordinates and observations in one inertial frame to those in another such frame.

luminiferous ether: A supposed medium for the transmission of light. The concept was rendered superfluous by the special theory of relativity early in the 20th century.

luminosity: The total power output of an object in the form of light. (Sometimes extended to include all forms of radiated energy.)

luminosity distance: The inferred distance to an astronomical object derived by comparing its observed brightness to its presumed total luminosity.

Mach's principle: The principle, elucidated by Ernst Mach, that the distribution of matter in the universe determines local inertial frames.

MACHO: Massive compact halo object. Any object such as a white dwarf, neutron star, or black hole that could account for some or all of the dark matter in the halos of galaxies.

magnetic monopole: A hypothetical particle representing one unit of magnetic "charge." Although required by grand unified and other theories, no magnetic monopole has been unequivocally observed.

main sequence: The curve on a Hertzsprung–Russell diagram along which stable hydrogen-fusing stars lie.

many-worlds interpretation: An interpretation of the measurement problem in quantum mechanics which holds that each act of measurement causes the universe to split into noncommunicating, parallel, quantum entities.

mass: That property of an object which causes it to resist changes in its state of motion; also, that property which generates gravitational attraction.

mass-to-light ratio: The ratio of the total mass of a luminous aggregate of matter expressed in solar masses to its total luminosity expressed in solar luminosities.

matter density parameter: The ratio of the average density in gravitating matter to the critical density, written Ω_M .

matter era: The epoch of the universe, lasting from approximately the time of recombination until the present, during which the energy density of radiation is negligible in determining the overall gravitational field of the universe, and the mass-energy density of matter is dominant.

measurement problem: The name for the enigma of how a measurement changes a quantum system into a definite state from one that evolves according to the probabilistic Schrödinger equation.

mechanics: The science of motion.

metal: In astronomy, all elements heavier than helium, regardless of whether they are chemically “metals” or not.

metric coefficient: The functions in the metric that multiply with the coordinate differentials (for example, Δx) to convert those differentials into physical distances.

metric equation: The expression that describes how to compute the distance between two infinitesimally separated points (or events) in a given geometry. Also called simply the “metric.”

microlensing: Gravitational lensing by relatively small objects such as stars or stellar remnants.

Milky Way: The name of our Galaxy. Also the name given to the band of diffuse light seen in the night sky that originates in the disk of our Galaxy.

Minkowskian space-time: The geometrically flat, four-dimensional space-time appropriate to special relativity.

model: A hypothesis or group of related hypotheses that describes and clarifies a natural phenomenon, entity, etc.

myth: A narrative intended to explain or justify the beliefs of a people. The term usually suggests a lack of historical and factual basis.

nebula: A cloud of gas or dust in space.

neutrino: Any of three species of very weakly interacting lepton with an extremely small mass.

neutron: A charge-neutral hadron that is one of the two particles that make up the nuclei of atoms. Neutrons are unstable outside the nucleus, but stable within it.

neutron degeneracy: A condition of matter in which electrons and protons are crushed together to form neutrons, and all quantum states available to the neutrons are filled.

neutron degeneracy pressure: A form of pressure that arises from neutron degeneracy, when the neutrons cannot be forced further together because of the exclusion principle.

neutron star: A dead “star” supported by neutron degeneracy pressure.

Newton's first law: The law of motion which states that an object in a state of uniform motion will remain in that state unless acted upon by an external force.

Newton's second law: The law of motion which states that the net applied force on an object produces an acceleration in proportion to the mass, $F = ma$.

Newton's third law: The law of motion which states that if A exerts a force on B, then B will exert an equal and oppositely directed force on A. For every action, there is an equal and opposite reaction.

Newtonian equivalence principle: The principle that the laws of mechanics are the same in inertial and freefalling frames of reference. This implies that gravitational mass and inertial mass are equivalent.

no-hair theorem: The theorem that the gravitational field of a black hole is entirely determined by only its mass, angular momentum, and any electric charge.

nova: An abrupt, very bright flare-up of a star. Most likely due to the accumulation of hydrogen from a companion upon the surface of a white dwarf. The pressure and temperature grow in this matter until a thermonuclear explosion is generated.

nuclear forces: Two of the fundamental forces, or interactions, the strong interaction and the weak interaction. Not necessarily confined exclusively to the nucleus, despite the name. The strong interaction not only holds nucleons together in the nucleus, but also binds quarks into hadrons. The weak interaction is involved in some nuclear processes such as radioactivity, but also causes free neutrons to decay.

nuclear reaction: A reaction that occurs in and may change the nucleus of at least one atom. Examples include radioactivity, fission, and fusion.

nucleon: Either of the two fermionic particles, the proton and the neutron, which form the nuclei of atoms.

nucleosynthesis: The process by which nuclear reactions produce the various elements of the Periodic Table.

nucleosynthesis epoch: The interval in the early history of the universe when helium was created, along with traces of a few other light-element isotopes.

nucleus: The central region of an atom, which gives it its elemental identity.

null: Of a space-time interval, capable of being traversed only by a massless particle such as a photon. A null or lightlike space-time interval is zero.

Occam's Razor : The principle that when all other things are equal, the simplest explanation is preferred.

Olbers' paradox: The fact that the night sky is dark even though in an infinite universe with stars that live forever, the night sky would be as bright as the surface of a star. The paradox disappears when it is realized that stars do not live forever and the universe is not infinitely old.

omega: See density parameter.

open universe: A standard model that expands forever and is infinite in space and time, although it begins with a big bang. Sometimes applied strictly to the hyperbolic standard model, although both the hyperbolic and flat models are open in the sense of expanding forever.

pair production: The creation of a particle and its antiparticle from some form of energy, such as photons.

parallax: The apparent shift in the position of a celestial object, such as a star, due to the changing vantage point of the observer. Astronomical parallax can be caused by phenomena such as the orbital motion of the Earth, or its daily rotation (*diurnal parallax*).

parameters of the universe: A set of measurable quantities that describe and distinguish the full set of homogeneous and isotropic models.

parsec (pc): That distance producing one second of arc of parallax over the baseline of the Earth's orbit. One parsec corresponds to 3.26 lightyears.

particle horizon: A surface beyond which we cannot see because the light from more distant objects has not had time to reach us over the age of the universe.

past: Those events that could have influenced a given event.

peculiar velocity: The unique velocity of an object such as a galaxy, due to its individual gravitational interactions with other objects and not due to the general cosmological recession.

perfect cosmological principle: The principle that the universe is unchanging, that is, homogeneous in time as well as in space. Refuted by the direct observation that the oldest objects in the universe are not like those in our immediate surroundings.

Periodic Table: A tabulation of the elements in increasing order of atomic number.

photon: A boson that is the particle of electromagnetic radiation (light). The photon is also the carrier particle of the electromagnetic force.

photon damping: The tendency of photons in the early universe to smooth out inhomogeneities in matter with which they are in thermal equilibrium.

photon sphere: The radius around a black hole at which light paths are gravitationally bent into a circle, thus causing the photons to orbit the hole.

Planck epoch: The epoch from the beginning of the universe until the Planck time. Very little is known about this interval, although probably all four fundamental forces were united.

Planck length: The Hubble length of the universe at the Planck time, approximately 10^{-33} cm.

Planck time: The cosmic time near the beginning of the universe, 10^{-43} s, at which classical gravity gained control of the universe as a whole.

Planck's constant: A fundamental constant of physics, h , which sets the scale of quantum mechanical effects.

plasma: A gas in which many or most of the atoms are ionized.

Population I, II, and III: Labels for the generations of stars, determined by the proportion of heavy elements contained in their members. Population I stars are youngest, while Population III represents the primordial stars.

positron: The antimatter partner of the electron.

potential: In physics, a mathematical function that describes the energy density of a field.

potential energy: The energy possessed by something by virtue of its location in a potential field, for example, its position in a gravitational field.

predictive power: The ability of a hypothesis or model to predict unobserved effects. This provides an important means of testing a hypothesis.

primordial element: One of those elements and isotopes formed in the big bang; specifically, hydrogen, helium (both helium-3 and helium-4), most deuterium and tritium, and some lithium-7.

principle of causality: The principle that a cause must always lie in the past of its effect for all possible observers.

principle of reciprocity: The principle in special relativity that two inertial frames will observe exactly the same phenomena when each observes the other. For example, each will see lengths in the other frame to be contracted by the same amount.

proper length: The length of an object measured in its own rest frame.

proper time: The time interval between two events as measured in the rest frame in which those events occurred. Numerically equal to the invariant space-time interval.

proton: A hadron which is one of the two particles that make up atomic nuclei. The proton is the least-massive baryon. Its absolute stability is uncertain, but its half-life is at least 10^{31} years.

pulsar: A rotating neutron star that emits regular, periodic bursts of radio emissions.

quantum: The smallest unit of some quantity.

quantum cosmology: A theory that attempts to describe the evolution of the universe in quantum mechanical terms.

quantum gravity: A unification of gravity and quantum field theory, not yet achieved.

quantum fluctuation: The small variation that must be present in a quantum field due to the uncertainty principle.

quantum mechanics: The theory that describes the behavior of the very small, such as molecules, atoms, and subatomic particles. Spectacularly successful at explaining experimental data, but gravity cannot yet be made to fit within the theory.

quantum state: A particular configuration of quantum properties, for example, energy, spin, momentum, charge, etc., that define a particular system.

quark: One of the six fundamental particles that make up hadrons.

quark epoch: The interval in the early universe during which quarks were unconfined in hadrons, and dominant.

quasar: An object that emits an extremely large luminosity from a small region. Invariably found only at large redshifts and hence distances. Also called *quasi-stellar objects* or *QSOs*.

quasi-stellar object (QSO): See quasar.

quintessence: A hypothetical exotic form of matter or energy that produces a negative pressure and a cosmological acceleration like a cosmological constant. A quintessence need not be constant in time, so the effective A force can change as the universe evolves.

radiation: The emission of particles or energy. Also the particle or energy so emitted.

radiation era: The epoch in the history of the universe, lasting from the big bang until approximately the time of recombination, during which the energy density of radiation controlled the gravity of the cosmos.

radioactive dating: The determination of the age of a sample by the measurement of the ratio of the decay products to the precursor, for one or more radioactive isotopes. Radioactive dating is possible because each unstable isotope has a well defined half-life.

radioactivity: Emission of particles from the nucleus of an atom.

reception distance: The distance of the source of light at the time the light was received.

recombination: The moment in the early universe when the temperature became sufficiently low that free electrons could no longer overcome the electrostatic attraction of the hydrogen nuclei and were captured to form atomic hydrogen. When this occurred the universe became transparent.

red dwarf: A small, dim, low-mass main sequence star.

red giant: A star near the end of its life; it fuses heavier elements in its core and has a greatly expanded outer layer.

redshift: A shift in the frequency of a photon toward lower energy.

redshift-distance relation: A theoretical relationship between the redshift of an object, such as a galaxy, and its distance from us. By measuring both distances and redshifts it is possible in principle to determine the evolution of the cosmic scale factor, $R(t)$.

regular cluster: A cluster of galaxies with a relatively smooth, approximately spherical shape. Most regular clusters are dominated by elliptical galaxies.

reionization: The point in time early in the universe, but after recombination, when the first stars formed and their ultraviolet light began to ionize the neutral hydrogen gas that filled the universe.

relativity: The rules relating observations in one inertial frame of reference to the observations of the same phenomenon in another inertial frame of reference. Casually applied only to the Einsteinian special theory of relativity, but actually a more general term.

relativity principle: The postulate of the special theory of relativity which states that the laws of physics are the same in all inertial frames of reference.

relevant: Of a scientific hypothesis: directly related to the phenomenon it seeks to explain.

relic problem: The unresolved issue in standard cosmology in which various theories of particle physics would invariably produce massive particles that are not observed.

rest energy: The energy corresponding to the rest mass according to $E = m_0c^2$.

rest mass: The mass of an object measured in its own rest frame. An important invariant quantity.

retrograde motion: The apparent reversal in the motion of a planet across the sky relative to the background stars, caused by the Earth passing the planet or being passed by it.

Riemannian geometry: A generalized geometry that has the property of being locally flat; that is, in a sufficiently small region, a Riemannian geometry can be approximated by a Euclidean or Minkowskian geometry.

Robertson–Walker metric: The metric that describes an isotropic and homogeneous cosmological space-time.

Sachs–Wolfe effect: The scattering of photons from perturbations in the early universe. Photons that last interacted with an overdense region suffer a gravitational redshift, whereas those that last scattered from an underdense region are blueshifted.

scale factor: The quantity (R) that describes how the scale changes in the expanding (or contracting) universe.

Schrödinger equation: The equation that describes the evolution of a nonrelativistic wavefunction.

Schwarzschild radius: The radius of the event horizon of a nonrotating black hole of mass M , equal to $2GM/c^2$.

scientific method: An investigative approach in which data are gathered, a hypothesis is formulated to explain the data, and further experiments are performed to test the hypothesis.

second law of thermodynamics: The law that states that the entropy of a closed system always increases or at best remains the same in any process.

seed perturbations: The initial small fluctuations in the universe that grow to become the observed cosmic background radiation temperature fluctuations, and eventually large-scale cosmic structure.

simplicity: The property of a scientific hypothesis that its proposed explanation must not be unnecessarily complicated.

simultaneity: The coincidence of the time coordinate of two events; the observation that two occurrences take place at the same time. Simultaneity is not invariant, but depends upon the reference frame of the observer.

singularity: In classical general relativity, a location at which physical quantities such as density become infinite.

solar luminosity (L_\odot): The energy output or luminosity of the Sun, used as a standard in astronomy.

solar mass (M_\odot): The mass of the Sun, used as a standard in astronomy.

spacelike: Of a space-time interval: incapable of being connected by anything that travels at or below the speed of light in vacuo.

space-time: The geometry that merges space and time coordinates.

space-time diagram: A depiction of space-time, usually including time and only one spatial dimension.

space-time interval: The invariant distance in space-time between two events, as specified by the metric equation.

spectrum: The components of emitted radiation, or a collection of waves separated and arranged in the order of some varying characteristic such as wavelength, frequency, mass, or energy.

speed: The magnitude of the velocity.

speed of light: The finite speed at which light travels. Unless otherwise stated, usually refers to the fundamental constant c , the speed of light in a perfect vacuum.

spherical geometry: A geometry that has positive constant curvature.

spiral galaxy: A galaxy that shows spiral arms, resembling a glowing pinwheel. Spirals typically contain a spheroidal nuclear bulge surrounded by a flat disk of stars, dust, and gas through which the spirals are threaded. The spirals themselves are delineated by bright young stars, and probably represent density waves traveling through the disk.

spontaneous symmetry breaking: The loss of symmetry that causes fundamental forces to become distinguishable. In most theories, this occurs in the early universe when the temperature becomes low enough that the different energy scales of the different forces become important.

standard candle: An object of known intrinsic luminosity, useful in the measurement of luminosity distances.

standard model: One of the set of big bang cosmological models derived with the minimum set of assumptions, namely that the cosmological principle holds and the cosmological constant is zero.

star: A self-luminous object held together by its own self-gravity. Often refers to those objects that generate energy from nuclear reactions occurring at their cores, but may also be applied to stellar remnants such as neutron stars.

static surface: The surface surrounding a Kerr black hole at which even light cannot resist being dragged along in the direction of the rotation of the hole.

steady state model: A cosmological model that obeys the perfect cosmological principle. Generally applied to specific models which contain a cosmological constant generated by the regular creation of matter.

stellar parallax: See parallax.

string theory: A theory in which the fundamental structure is not a pointlike particle but is a quantum string, whose vibrations are associated with what we call particles.

strong equivalence principle: See Einstein equivalence principle

strong interaction: The fundamental force that binds quarks into hadrons and holds nucleons together in atomic nuclei. Sometimes called the strong force or the strong nuclear force.

structure formation: The process by which the large-scale structure in the universe, namely the galaxies, galaxy clusters, and superclusters, developed from small density perturbations in the early universe.

structure problem: The incompletely resolved difficulty of explaining the origin of structure, representing local inhomogeneities, in a universe that is isotropic and homogeneous on the largest scales.

supercluster: A cluster of galaxy clusters.

supernova: The explosive death of a star. Type Ia supernovae probably occur when a white dwarf accumulates upon its surface too much gas from a companion, causing the white dwarf to exceed the Chandrasekhar limit. Type II supernovae occur when a massive star has reached the end point of nuclear fusion and can no longer support itself. In both cases, the result is a catastrophic gravitational collapse and an explosion so violent that elements heavier than iron are created. Any remaining core becomes a neutron star or a black hole.

superposition of states: In quantum mechanics, the description of an unobserved system in terms of the probabilities of all possible states.

surface of last scattering: The point at recombination at which the cosmic background photons last interacted with the baryonic matter. After this, the cosmic background photons streamed freely through space. The surface of last scattering is what is seen when the cosmic background radiation is observed.

symmetry: The property under which some quantity does not change when certain attributes, such as spatial location, time, rotation, and so forth, vary.

temperature: A measure of the average kinetic energy of random motion of the constituents (for example, molecules, atoms, or photons) of a system.

testable: Of a hypothesis, capable of being tested because it makes a specific prediction. Similar to *falsifiable*.

theory: In scientific usage, a hypothesis or related group of hypotheses that have become well established.

thermal equilibrium: A state in which energy is equally distributed among all particles, and all the statistical properties of the particles can be described by a single parameter, the temperature.

thermal radiation: Radiation emitted by any object with a temperature greater than absolute zero. A thermal spectrum occurs because some of the heat energy of the object is converted into photons. In general, a thermal spectrum depends not only upon the temperature, but also upon the composition of the object, its shape, its heat capacity, and so forth. Compare *blackbody radiation*.

thermodynamics: The theory of heat and its relationship to other forms of energy.

thought experiment: An experiment that could be performed in principle but might be very difficult in practice, and whose outcome can be predicted by pure logic. Often used to develop the consequences of a theory, so that more practical phenomena can be predicted and put to actual experimental tests.

tidal force: In Newtonian gravity, the net force on an extended body due to a difference in gravitational force from one region of a body to another. In general relativity, a force arising when nearby geodesics diverge in space-time, because the worldlines of all parts of an extended body cannot travel along a single geodesic.

time dilation: An apparent decrease in the rate of the flow of time (for example, the ticking of a clock) in a frame moving relative to a given observer, determined by the Lorentz transformation from one frame to the other.

timelike: Of a space-time interval: capable of being connected by anything that travels below the speed of light in vacuo. Worldlines of physical objects follow timelike paths through space-time.

top-down structure formation: The formation of large structures, such as galaxy superclusters or perhaps even the vast filaments and voids, prior to the formation of smaller structures such as individual galaxies.

true vacuum: A stable state in which a quantum field is zero and the corresponding potential is also zero; that is, the vacuum energy density is zero.

Tully–Fisher relationship: An empirical relationship between the width of the 21-cm line of hydrogen emissions from spiral galaxies and the mass of the galaxy. The relationship arises because a larger mass increases the rotation rate, and a faster rotation causes a broader line; the precise calibration must be determined observationally.

turnoff mass: The mass of the largest star in a cluster that is still on the main sequence. The age at which a star moves from the main sequence to the red giant phase depends almost entirely upon its mass and chemical composition, with more massive stars leaving the main sequence earlier. The stars in a cluster all formed at essentially the same time and have similar chemical composition, so the turnoff mass can be used to determine the age of the cluster.

uncertainty principle: The principle of quantum mechanics which states that the values of both members of certain pairs of variables, such as position and momentum, or energy and time interval, cannot be determined simultaneously to arbitrary precision. For example, the more precisely the momentum of a particle is measured, the less determined is its position. The uncertainty in the values of energy and time interval permits the quantum creation of virtual particles from the vacuum.

unified epoch: That interval in the early history of the universe when three of the four fundamental forces, the strong and weak interactions and the electromagnetic force, were unified.

uniform motion: Motion at a constant velocity. The state of rest is a special case of uniform motion.

universal gravitation: Newton's mathematical formulation of the law of attraction between two masses: $F_g = GM_1 M_2 / R^2$.

universe: That which contains and subsumes all the laws of nature, and everything subject to those laws; the sum of all that exists physically, including matter, energy, physical laws, space, and time.

vacuum energy: The energy associated with empty space, that is, the vacuum itself.

vector: A mathematical entity that has direction as well as magnitude. Important physical quantities represented by vectors include velocity, acceleration, and force. A vector changes whenever either its direction or its magnitude changes.

velocity: The rate of change of displacement with time. Velocity includes both the speed of motion and the direction of motion.

Virgo Cluster: A nearby irregular cluster of galaxies located in the constellation Virgo. The distance to the Virgo cluster is an important rung in the distance ladder.

virial theorem: A statistical result that relates the mean gravitational field of a cluster to the dispersion of the velocities of the members of the cluster.

virtual particles: Particles that exist only as permitted by the uncertainty principle.

void: In astronomy, a huge region of space that is unusually empty of galaxies. Recent research has shown that voids are not entirely empty, but they are underdense and contain far fewer bright galaxies than average.

wave: A propagating disturbance that transmits energy from one point to another without physically transporting the oscillating quantity.

wave amplitude: The size of the departure from the average of the quantity that supports the wave.

wave frequency: The number of wave crests that pass a fixed point in a fixed interval of time.

wavefunction: The quantity that obeys the Schrödinger equation. In the Copenhagen interpretation of quantum mechanics, the wavefunction is a mathematical entity that describes the probabilities that the quantum system will assume any of several possible states upon a measurement.

wavefunction of the universe: A wavefunction that treats the scale factor as a quantum variable and describes its evolution in quantum, rather than classical general relativistic, terms.

wavelength: The distance from one crest of a wave to the next.

weak equivalence principle: *See* Newtonian equivalence principle.

weak interaction: The fundamental force that accounts for some particle interactions, such as beta

decay, the decay of free neutrons, neutrino interactions, and so forth. Sometimes called the weak force or the weak nuclear force.

weight: The gravitational force experienced by an object. Usually refers to the gravitational attraction due to a large object, such as a planet, upon smaller objects at or near its surface.

white dwarf: A compact stellar remnant supported by electron degeneracy pressure and shining only by the diffusion of light from its interior. White dwarfs cool slowly; if the universe exists long enough they will all cool into nonluminous black dwarfs.

WIMP: *Weakly interacting massive particle.* A particle with a nonzero mass which participates only in the weak interaction.

work: In physics, a compound of the force exerted with the displacement produced.

worldline: The path of a particle in space-time.

This page intentionally left blank

Bibliography

Mythology, Philosophy, and History References

- Beier, Ulli, 1966. *The Origin of Life and Death: African Creation Myths*. London: Heinemann Educational Books Ltd.
- Birch, Cyril, 1961. *Chinese Myths and Fantasies*. New York: Henry Z. Walck, Inc.
- Brundage, Burr Cartwright, 1979. *The Fifth Sun: Aztec Gods, Aztec World*. Austin: University of Texas Press.
- Cambell, Joseph, 1988. *The Power of Myth*. New York: Doubleday.
- Copi, Irving M., 1972. *Introduction to Logic*, Fourth Edition. New York: Macmillan Publishing Company Inc.
- Gingerich, Owen, and J. R. Voelkel, 1998. "Tycho Brahe's Copernican Campaign." *J. History of Astronomy*, 29, 1.
- Hetherington, Noriss S., editor, 1993. *Cosmology: Historical, Literary, Philosophical, Religious, and Scientific Perspectives*. New York: Garland Publishing, Inc.
- Koestler, Arthur, 1968. *The Sleepwalkers*. New York: Macmillan.
- Kragh, Helge, 1996. *Cosmology and Controversy*. Princeton: Princeton University Press.
- Pais, Abraham, 1982. "Subtle Is the Lord": *The Science and the Life of Albert Einstein*. Oxford: Oxford University Press.
- Popper, Karl, 1959. *The Logic of Scientific Discovery*. London: Hutchinson.
- Taylor, Colin F., editor, 1994. *Native American Myths and Legends*. New York: Smithmark Publishers Inc.
- Tuchmann, Barbara W., 1984. *The March of Folly*. New York: Ballantine Books.
- Walls, Jan and Yvonne, 1984. *Classical Chinese Myths*. Hong Kong: Joint Publishing Company.

Popular-Level Books on Cosmology and Related Subjects

- Barrow, John, 1994. *The Origin of the Universe*. New York: Basic Books.

- Cosmology + 1: Readings from Scientific American.* San Francisco: W. H. Freeman, 1977.
- Ferris, Timothy, 1988. *Coming of Age in the Milky Way.* New York: William Morrow and Company, Inc.
- Greene, Brian, 1999. *The Elegant Universe.* New York: W. W. Norton.
- Hawking, Stephen, 1988. *A Brief History of Time.* Toronto: Bantam Books.
- Penrose, Roger, 1989. *The Emperor's New Mind.* Oxford: Oxford University Press.
- Silk, Joseph, 1989. *The Big Bang.* Revised and Updated Edition. New York: W. H. Freeman and Company.
- Silk, Joseph, 1994. *A Short History of the Universe.* Scientific American Library. New York: W. H. Freeman and Company.
- Thorne, Kip S., 1994. *Black Holes and Time Warps: Einstein's Outrageous Legacy.* New York: W.W. Norton.
- Weinberg, Steven, 1988. *The First Three Minutes.* Updated Edition. New York: Basic Books.
- Weinberg, Steven, 1992. *Dreams of a Final Theory.* New York: Pantheon Books.
- Zuckerman, Ben, and Matthew Malkan, 1996. *The Origin and Evolution of the Universe.* Boston: Jones and Bartlett.

Magazine Articles

- Albert, David Z., 1994. "Bohm's Alternative to Quantum Mechanics." *Scientific American*, 270, No. 5, 58.
- Brashear, Ronald S., Donald E. Osterbrock, and Joel A. Gwinn, 1993. "Edwin Hubble and the Expanding Universe." *Scientific American*, 269, No. 1.
- Deutsch, David, and Michael Lockwood, 1994. "The Quantum Physics of Time Travel." *Scientific American*, 270, No. 3, 68.
- DeWitt, Bryce, 1983. "Quantum Cosmology." *Scientific American*, 249, No. 6, 112.
- Halliwell, Jonathan J., 1991. "Quantum Cosmology and the Creation of the Universe." *Scientific American*, 265, No. 6, 76.
- Gould, Stephen Jay, 1994. "The Evolution of Life on Earth." *Scientific American*, 271, No. 4, 84–91.
- Kirschner, Robert, 1994. "The Earth's Elements." *Scientific American*, 271, No. 4, 58–67.
- Peebles, P. James, *et al.*, 1994. "The Evolution of the Universe." *Scientific American*, 271, No. 4, 52–57.
- Perlmutter, S., 2003. "Supernovae, Dark Energy, and the Accelerating Universe." *Physics Today*, 56, 53.
- Schilling, Govert, 2003. "Cosmology's Treasure." *Sky and Telescope*, 105, 32.
- Schramm, David N., 1994. "Dark Matter and the Origin of Cosmic Structure." *Sky and Telescope*, 88, 28–35.
- Schwarzschild, B., 2003. *Physics Today*, 56, 21.
- Schwarzschild, B., 2003. *Physics Today*, 57, 19.
- Weinberg, Steven, 1994. "Life in the Universe." *Scientific American*, 271, No. 4, 44–51.

Introductory Texts on Astronomy, Cosmology, and Physics

There are dozens of excellent introductory texts on general astronomy and physics. We list here only a few.

- Chaisson, Eric, and Steve McMillan, 2002. *Astronomy Today*, Fourth Edition. Upper Saddle River, New Jersey: Prentice Hall.
- Harrison, Edward R., 2000. *Cosmology: The Science of the Universe*, Second Edition. Cambridge: Cambridge University Press.
- Kaufmann, William J., and Roger A. Freedman, 1998. *Universe*, Fifth Edition. New York: W. H. Freeman and Company.
- Kuhn, Karl F., 1994. *In Quest of the Universe*, Second Edition. Minneapolis: West Publishing Company.
- Resnick, Robert, David Halliday, and Kenneth S. Krane, 1992. *Physics*, Fourth Edition. New York: John Wiley and Sons Inc.

Advanced Books and Texts

- Adler, Ronald, Maurice Bazin, and Menahem Schiffer, 1975. *Introduction to General Relativity*, Second Edition. New York: McGraw-Hill Book Company.
- Allen, C. W., 1976. *Astrophysical Quantities*, Third Edition. Dover: Athlone Press.
- Bernstein, Jeremy, 1995. *An Introduction to Cosmology*. Englewood Cliffs, New Jersey: Prentice Hall.
- Einstein, Albert, *et al.*, 1952. *The Principle of Relativity*. New York: Dover.
- Hawking, Stephen, and W. Israel, editors, 1979. *General Relativity: An Einstein Centenary Survey*. New York: Cambridge University Press.
- Kolb, Edward W., and Turner, Michael S., 1990. *The Early Universe*. Redwood City: Addison Wesley Publishing Company.
- Mandolesi, N., and Vittorio, N., editors , 1990. *The Cosmic Microwave Background: 25 Years Later*. Dordrecht: Kluwer Academic.
- Merzbacher, Eugen, 1970. *Quantum Mechanics*, Second Edition. New York: John Wiley & Sons, Inc.
- Metcalf, Nigel, and Shanks, Tom, editors, 2002. *A New Era in Cosmology*. San Francisco: Astronomical Society of the Pacific.
- Misner, Charles W., Kip S. Thorne, and John Archibald Wheeler, 1973. *Gravitation*. San Francisco: W. H. Freeman and Co.
- Peacock, John A., 1999. *Cosmological Physics*. Cambridge: Cambridge University Press.
- Peebles, P. J. E., 1993. *Principles of Physical Cosmology*. Princeton, New Jersey: Princeton University Press.
- Perkins, Donald H., 1982. *Introduction to High Energy Physics*, Second Edition. Reading, Massachusetts: Addison-Wesley.

- Rees, M., 1995. *Perspectives in Physical Cosmology*. Cambridge: Cambridge University Press.
- Rindler, Wolfgang, 1977. *Essential Relativity*, Second Edition. New York: Springer-Verlag.
- Rowan-Robertson, Michael, 1977. *Cosmology*. Oxford: University of Oxford Press.
- Schwarzschild, Martin, 1958. *Structure and Evolution of the Stars*. New York: Dover Publications Inc., reprinted from Princeton University Press.
- Shapiro, Stuart L, and Saul A. Teukolsky, 1983. *Black Holes, White Dwarfs, and Neutron Stars: The Physics of Compact Objects*. New York: John Wiley & Sons.
- Vangioni-Flam, E., M. Casse, J. Audouze, and J. Tran Thuanh Van., editors, 1990. *Astrophysical Ages and Dating Methods*. Gif sur Yvette, France : Editions Frontieres.
- Weinberg, Steven, 1972. *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity*. New York: John Wiley and Sons.

Technical Journal Articles

- Bennett, C. L., *et al.* 2003. "First-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Preliminary Maps and Basic Results." *Astrophys. J. Supp.*, 148, 1.
- Bolte, M., and C. J. Hogan, 1995. "Conflict Over the Age of the Universe." *Nature*, 376, 399.
- Coles, P., and G. Ellis, 1994. "The Case for an Open Universe." *Nature*, 370, 609.
- Coulson, D., P. Lerreira, P. Graham, and N. Turok, 1994. "Microwave Anisotropies from Cosmic Defects." *Nature*, 368, 27.
- Dunlop, J. S. *et al.* 1994. "Detection of a Large Mass of Dust in a Radio Galaxy at $z = 3.8$." *Nature*, 370, 347.
- Elston, R., K. L. Thompson, and G. J. Hill, 1994. "Detection of strong iron emission from quasars at redshift $z > 3$." *Nature*, 367, 250.
- Freedman, W. L., *et al.* 1994, "Distance to the Virgo Cluster Galaxy M100 from Hubble Space Telescope Observations of Cepheids." *Nature*, 371, 757.
- Fukugita, M., C. J. Hogan, and P. J. E. Peebles, 1996. "The History of the Galaxies." *Nature*, 381, 489.
- Gott, R., *et al.*, 1974. "An Unbound Universe." *Astrophys. J.*, 194, 543.
- Hogg, D. W. 2000. "Distance Measures in Cosmology." astro-ph/9905116.
- Izotov, Y. I. *et al.*, 1999. "Helium Abundance in the Most Metal-Deficient Blue Compact Galaxies: I Zw 18 and SBS 0335-052." *Astrophys. J.*, 527, 757.
- Johnson, H. L., and W. W. Morgan, 1953. "Fundamental Stellar Photometry for Standards of Spectral Type on the Revised System of the Yerkes Spectral Atlas." *Astrophys. J.*, 117, 313.

- Johnson, H. L., and A. R. Sandage, 1956. "Three Color Photometry in the Globular Cluster M3." *Astrophys. J.*, 124, 379.
- Lin, H., *et al.*, 1996. "The Power Spectrum of Galaxy Clustering in the Las Campanas Redshift Survey." *Astrophys. J.*, 471, 617.
- Livio, M., 2003. "The World According to the Hubble Space Telescope." astro-ph/0303500.
- Maller, A. H., D. H. McIntosh, N. Katz, and M. D. Weinberg, 2003. "The Clustering Dipole of the Local Universe from the Two Micron All Sky Survey." *Astrophys. J.*, 598, L1.
- Mather, J. C., *et al.*, 1990. "A Preliminary Measurement of the Cosmic Microwave Background Radiation Spectrum by the Cosmic Background Explorer Satellite." *Astrophys. J.*, 354, L37.
- Peacock, J. A., 2003. "Cosmological Parameters from the Microwave Background and Large-Scale Structure." In *Proceedings of the Fourth International Workshop on the Identification of Dark Matter*, N. J. C. Spooner, and V. Kudryavtsev, editors. Singapore: World Scientific, 1.
- Perlmutter, S., *et al.*, 1999. "Measurements of Ω and Λ from 42 High-Redshift Supernovae." *Astrophys. J. Supp.*, 517, 565.
- Pierce, M. J., *et al.*, 1994. "The Hubble Constant and Virgo Cluster Distance from Observations of Cepheid Variables." *Nature*, 371, 385.
- Roth, K. C., D. M. Meyer, and I. Hawkins, 1993. "Interstellar Cyanogen and the Temperature of the Cosmic Microwave Background Radiation." *Astrophys. J.*, 413, L67.
- Songaila, A., *et al.*, 1994. "Measurement of Microwave Background Temperature at a Redshift $z = 1.776$." *Nature*, 371, 43.
- Strauss, M. A., *et al.*, 1992. "A Redshift Survey of IRAS Galaxies. IV. The Galaxy Distribution and the Inferred Density Field." *Astrophys. J.*, 385, 421.
- Tanaka, Y., *et al.*, 1995. "Gravitationally Redshifted Emission Implying an Accretion Disk and Massive Black Hole in the Active Galaxy MCG-6-30-15." *Nature* 375, 659.
- Weisberg, J. M., and J. H. Taylor, 2003. "The Relativistic Binary Pulsar PSR 1913+16." in *Radio Pulsars*, M. Bailes, D. J. Nice, and S. E. Thorsett, editors. San Francisco: Astronomical Society of the Pacific.
- Wilson, T. L., and R. T. Rood, 1994. "Abundances in the Interstellar Medium." *Ann. Rev. Astron. Astrophys.*, 32, 191.

This page intentionally left blank

INDEX

- 2MASS
see Two-Micron All Sky Survey
- 3C273 (quasar) 268
- 61 Cygni (star) 285
- Abell 1689 (galaxy cluster) 436
- Abell 2218 (galaxy cluster) 440
- aberration of starlight 46
- absolute zero 98, 104, 502, 516, 521
- absorption spectrum 107, 521
- acceleration 61–2, 63, 80, 166–7, 179, 223, 521
 and gravity 216–19
- accretion disk 138, 148, 265, 273, 521
 black holes in 268–70
- active galaxies 268, 269, 270, 271, 273, 521
- Adams, John 72
- Adams, W. S. 411
- AGN
see active galactic nuclei
- air resistance 32
- Airy, George 72
- Albrecht, Andreas 477
- Almagest* (Ptolemy) 12, 35
- Alpha Centauri 45, 111
- Alpher, Ralph 339, 340, 363, 411
- amplitude
see wave amplitude
- Andromeda Galaxy 113, 287, 288, 289, 435
- angular measure 517
- angular power spectrum 428, 429
- angular size 389, 404, 428–9, 521
- angular size-distance relationship 389–90, 391
- animism 6
- anthropic principle 157, 179, 480–1, 521
 strong 157–8
 weak 157
- anthropocentrism 12–13, 22, 159
- anthropomorphism 6, 22, 521
- antimatter 94, 115, 350, 521
- antiparticle 94
- Aquinas, Thomas 37
- Aristarchus 33–4, 35
- Aristotelian cosmology 29–32, 38, 43–4, 52, 163–4
 flaws in 32–3
- Aristotelian mechanics 29–30, 32–3, 52–3, 61
- arrow of time 373–4, 377, 511, 521
 and entropy 502
- astrolabe 38
- astrology 12–13
- astronomical constants 519
- astronomical unit 46, 48, 285, 516, 521
- astronomy 5, 22, 38, 108, 521
- atom 86–89, 94, 521
- atomic clocks 190, 222
- atomic structure 87–9, 90
- Baade, Walter 141, 343
- Babylon 7–8
- Bacon, Sir Francis 17
- bacteria 76, 78
- baryogenesis 357, 377, 521
- baryonic mass density 399
- baryon number
 conservation 357, 521
- baryons 356–7, 365, 366–7, 377, 381, 441–2, 459, 521
- Becquerel, Henri 90
- Bekenstein, Jacob 262
- Bell, Jocelyn 141
- Bessel, Friedrich Wilhelm 45, 285
- beta decay 92
- Beta Pictoris 124–5
- Bethe, Hans 77
- biased galaxy formation 453–4, 461, 522
- big bang 5, 6–7, 21, 156, 164, 319–20, 330–1, 334, 337, 339–40, 344, 379, 412, 488, 522
 and cosmic background radiation 416
 limitations 463–4
 nucleosynthesis 339–41, 342
- big crunch 320, 334, 371, 374, 522
- binary pulsars 239
- binary systems 138
 black holes in 264–5, 266
 neutron stars in 144–5, 146
- Birkhoff's theorem 255, 272, 521
- blackbody 104, 115, 416, 489, 522
see also cosmic background radiation (CBR)
- blackbody radiation 104–6, 116, 261, 409, 413, 522
- black holes 146, 147, 245–6, 271–2, 502, 522
 density 254
 detecting 263–5
 gravitational field 254–6

- black holes (*cont.*)

in active galactic nuclei 268–9

lightcones around 251–2

rotating 256–9, 270–1

singularity 253–4, 488

temperature 261–2

tidal forces around 250–1

black hole thermodynamics 262, 272, 502–3, 522

Bl Lacertae objects (elliptical galaxies) 268

blueshift 101–2, 114–15, 116, 278, 279, 309, 522

Bohm, David 491

Bohr, Niels 88, 93, 492

Boltzmann constant 128

Bólyai, János 226

Bondi, Hermann 338

boost factor 188, 209, 522

of Concorde flight 190

bosons 93–4, 115, 446, 522

bottom-up structure formation 437, 460, 522

Bradley, James 46

brown dwarfs 123, 124, 147, 396–7, 522

Brownian motion 177

Bunsen, Robert 107

Burbidge, Geoffrey 341

Burbidge, Margaret 341

 carbon 157

carrier bosons 94–5, 446, 522

Casimir effect 472, 473

catastrophism 74

causality, principle of

see principle of causality

Cavendish, Henry 68

CBR

see cosmic background radiation

CCDs

see charge-coupled devices

CDM

see cold dark matter

celestial spheres 27–8

central force 70

centrifugal force 169, 215

centripetal force 215

 Cepheid variables 287–8, 289, 291, 309, 385, 522

in Virgo Cluster 382–3

Chandrasekhar limit 139, 140, 148, 522

chaotic inflation 480–1, 484, 522

charge 86–87, 93, 257, 522

charge-coupled devices (CCDs) 352, 452

circular motion 33, 49, 63, 167

Aristotelian 30–1, 52, 163–4

Copernican 41

Clausius, Rudolf 97

closed universe 325, 334, 522

CNO cycle 130–1

COBE

see Cosmic Background Explorer

cold dark matter (CDM) 399, 404, 445–7, 461, 522

cold dark matter (CDM) model 458, 522

collapse of the wavefunction 492–3, 496, 511, 522

collisionless damping 448, 461, 522

Columbus, Christopher 38

Coma Berenices (constellation) 436

comets 71

comoving coordinates 299–300, 310, 522

Compton wavelength 351–2, 376, 522

concordance model 403, 404, 428–9, 431, 522–3

conservation of angular momentum 142–3, 148, 523

conservation of baryon number 357, 521

conservation of energy 97, 116, 415, 523

conservation of matter 97, 116, 523

conservation of momentum 62–3, 80, 523

 consistency (of scientific theory) 19, 523

continuum 103

coordinates 165, 179, 523

coordinate singularity 250, 272, 523

coordinate system 165–6

Copenhagen interpretation 492–4, 505, 510, 511, 523

Copernican principle 41, 55, 78, 85, 159, 161, 179, 523

Copernican revolution 39–42, 53, 55, 523

Copernicus, Nicholas 39, 78, 288

corpuscular theory of light 58, 102, 351, 489

Coriolis force 169–70, 179

correlation function 450, 461, 523

three-point 451

two-point 450–1

Cosmic Background Explorer (COBE) 382, 418, 419, 420, 427, 430

cosmic background

radiation (CBR) 340, 344–5, 347, 430, 431, 523

and cosmological principle 416–17

anisotropy 418–20, 422–3

blackbody 417–18

discovery 409–12

temperature-redshift relationship 413

cosmic censorship hypothesis 253, 272, 523

cosmic distance ladder 284–7, 309, 523

cosmic dust 120, 287

cosmic rays 93

cosmic rest frame 420–1

cosmic time 302, 310, 499, 523

cosmological constant 296–8, 309–10, 326–8, 334, 401, 476, 523

- models with 327, 332–3
 cosmological models 156,
 163, 298, 313–14,
 318–19, 331, 333–4,
 337–8, 379–80, 403
 limitations 463–4
 cosmological principle
 162–3, 179,
 343, 523
 and cosmic background
 radiation 416–17
 see also homogeneity;
 isotropy; perfect
 cosmological
 principle
 cosmological redshift 280,
 294–5, 305–7, 309,
 523
 cosmological simulation
 456–9
 cosmology 4–5, 21, 22, 523
 application of general
 relativity 295–7
 Greek 25–8
 observational programs
 381–2
 quantum 496–8, 511, 531
 scientific 13, 16
 timeline for discoveries
 39
 see also mythological
 cosmology
 cosmos
 see universe
 Coulomb barrier 129
 covalent bond 89
 Crab Nebula 142
 critical density 323, 334,
 441, 523
 crucial experiment 19, 523
 crystals 13–14
 Curtis, Heber 283, 284, 287
 curvature constant 301,
 310, 523
 Cygnus A (galaxy) 266–7,
 268
 Cygnus X-1 (binary
 system) 264–5
- dark age 437, 460, 523
 dark energy 296–7, 310,
 327, 400–2,
 405, 523
 dark halos 395, 398, 404,
 441, 523
 dark matter 147, 381, 398,
 404, 439–40, 461, 524
 and structure formation
 448–9
 cold 399, 404, 445–7, 522
 hot 399, 404, 445, 526
 darkness at night 407–9
 Darwin, Charles 75, 78
 data 18, 22, 54, 524
 dating techniques 78–9
 Davy, Sir Humphrey 19
 deceleration parameter
 324–5, 334,
 381, 524
 deduction 17
 deferents 36, 40
 deism 73
 density parameter 333,
 366, 388, 524
 see also matter density
 parameter,
 lambda density
 parameter
 de Sitter effect 298
 de Sitter model 298, 310,
 329, 334, 476, 497–8,
 524
 de Sitter, Willem 178, 297
 deuterium 129–30, 147,
 362–3, 377, 524
 abundance 364–5
 deuterons 129, 362
 De Witt, Bryce 496
*Dialogues Concerning Two
 Chief World Systems*
 (Galileo) 53
 Dicke, Robert 219, 410–11
 Dirac, P. A. M. 470, 488
 Dirac theory 488
 dispersion 103
 distance 191
 and angular size 389–90,
 391
 and redshift 289–90,
 383–4
 emission 383, 404
 interstellar 284–7
 luminosity 109–10,
 286–7, 309, 384
 reception 383–4, 404
 distance ladder
 see cosmic distance
 ladder
 Doppler effect 100–1, 116,
 278, 524
 relativistic 194
 Doppler shift 101, 145, 270,
 278–9, 306–7
 relative to cosmic
 background
 radiation 420
 Doroshkevich, A. G. 411
 dust grains 121–2
 dynamical methods 394,
 399, 404, 524
 dwarf irregular galaxies 112
- Earth
 age 77, 91
 distance from the Sun 34
 motion 421
 stationary 26–7, 31
 Eddington, Arthur, 77, 237,
 337
 Eddington model 337
 Einstein, Albert 19, 93,
 102, 106, 176–7, 178,
 179, 184, 232, 235,
 295, 401, 489, 496
 Einstein-de Sitter model
 325, 334,
 347–8, 524
 Einstein equivalence
 principle 219–21,
 242, 524
 Einstein's equations 235–6,
 257, 271, 313–14,
 504–5
 El Caracol temple (Mexico)
 10
 electromagnetic fields 96
 electromagnetic force 94,
 95, 116, 524
 electromagnetic spectrum
 102–3, 277, 524
 electromagnetic waves
 102–3
 electromagnetism 174–5,
 177–8, 182, 191
 electron degeneracy 136–7,
 146, 147,
 148, 524
 electron degeneracy
 pressure 139, 524
 electron microscopes 490
 electron neutrinos 443, 444
 electrons 86–7, 89, 90, 107,
 524
 electroweak interactions
 359, 377, 524

- elementary particles 86, 356–7
 elements 86, 87, 439, 524
 ellipse 48, 49, 55, 524
 elliptical galaxies 112, 268, 289, 524
 formation 439
 mass-to-light ratio 395–6
 elsewhere 201, 202, 524
 emission distance 383, 404, 524
 emission spectrum 107, 524
 energy 96, 114, 116, 372, 414–15, 524
 and entropy 501–2
 and matter 348–50
 conservation of 97, 116, 415
 relativistic 193–4, 196
 unit 517
 energy density 313, 345–6, 413
 energy flux 109
 entropy 98, 116, 371–4, 376, 377, 511, 524
 and arrow of time 502
 and energy 501–2
 and second law of thermodynamics 500–1
 of universe 502–3
Enuma Elish 7, 8
 Eötvös, Roland von 219
 epicycle 36
 equal density epoch 369, 377, 524
 equation of state 346, 401
 equilibrium 116, 524
 equivalence principle 52–3, 54, 524
 Einstein 219–21, 242
 Newtonian 217–19, 242
 Eratosthenes 26
 ergosphere 258–9, 272, 524
 escape velocity 245, 272, 316, 525
 ether 174–5, 185, 180, 528
 rejection of 183–4
 ether drag 183, 193
 Euclidean geometry 225
 Eudoxus 27, 28, 31
 eukaryotes 78
 event horizon 249–50, 258, 272, 465, 484, 525
 events 185, 209, 525
 simultaneous 201
 evolution (biology) 75–7
 exclusion principle 94, 115, 525
 expectation value 491
 experiment 16, 23, 525
 explanatory power 20, 22, 525
 extinction 109–10, 287, 309, 525
 extrasolar planets 125
 false vacuum 475, 480, 484, 525
 falsifiability (of scientific theory) 22, 525
 Fermi, Enrico 91, 92
 fermions 93–4, 95, 115, 525
 fictitious forces 169, 215
 Field, George 411
 fields 96, 352, 525
 “finger of God” phenomenon 452–3
 first acoustic peak 426, 428, 431, 525
 Fisher, J. Richard 292
 fission
 see nuclear fission
 Fitzgerald contraction hypothesis 183, 184
 Fitzgerald, George F. 183
 Flamsteed, John 72
 flatness problem 468–9, 478, 483, 484, 525
 flat geometry 226, 242, 298, 301, 302, 525
 flat universe 325–6, 334, 347, 525
 flux 525
 Follin, James 340
 force 30, 53, 55, 60, 61, 63, 80, 525
 unit 517
 Foucault pendulum 170, 223–4
 Fowler, William 341
 frame dragging 257–8
 frames of reference 167–8, 179, 182, 215, 525
 inertial 168, 170–1, 173, 179, 184–5, 200, 203–4
 Franklin, Benjamin 86
 Fraunhofer lines 107
 Freedman, Wendy 383
 freefall 170–1, 215–17, 224, 234, 242, 525
 frequency
 see wave frequency
 friction 61
 Friedmann, Alexander 178, 318
 Friedmann equations 318–19, 320–2, 323, 325, 327–8, 334, 345, 496, 525
 Frisch, Otto 91
 fundamental forces 94, 525
 see also electromagnetic force, weak nuclear force, strong nuclear force, gravity
 fundamental peak 426
 fundamental scale 426–7
 fusion
 see nuclear fusion
 future 201, 202, 209, 525
 galactic astronomy 289
 galactic cannibalism 439, 461, 525
 galactic motions 422
 galaxies 111, 113–14, 115, 116, 435–6, 459, 525
 active 268
 discovery of other 280, 288–9
 elliptical 112, 268, 289, 395–6
 formation 436–9
 biased 453–4, 461, 522
 irregular 112, 436
 nonluminous matter 396
 number counts 391–2
 radio 268
 spiral 111–12, 289, 395
 galaxy clusters 113, 116, 435, 460, 525
 quantitative measure 449–51
 studies and surveys 451–2, 454–6
 galaxy redshift survey 451
 Galilean moons 52
 Galilean relativity 173, 179, 180, 181–2, 184, 525

- Galilean transformation 173, 174, 182, 192
 Galileo Galilei 50, 53–4, 60, 67, 219
 Gamow, George 77, 339, 343, 363, 411
 gauge bosons
 see carrier bosons
 Gauss, Carl Friedrich 226
 general theory of relativity 178, 213–14, 231–2, 241
 application to cosmology 295–7
 mathematical 233, 235–7
 tests 237–41
 genetics 76
 geocentric model 26–8, 55, 159, 525
 geodesics 231–2, 242, 525
 geology 73
 geometry 225, 228–9
 Euclidean 225–6, 390–1, 525
 non-Euclidean 226
 of universe 389–92, 403
 Gliese 229 (star) 123, 124, 397
 Global Positioning Systems (GPS) 222
 globular clusters 78, 134–5, 147, 283–4, 526
 and age of universe 388
 gluons 95, 358, 377, 526
 Gödel, Kurt 504
 Gödel solution 504–5
 Gold, Thomas 141, 338
 GPS
 see Global Positioning Systems
 grandfather paradox 504, 506, 511, 526
 grand unified theories (GUTs) 356–8, 376, 470, 526
 gravimeters 68
 gravitational constant 68, 79, 80, 526
 gravitational field 96
 of black holes 254–6
 gravitational lens 256, 271, 272, 526
 gravitational lensing 256, 440, 441
 gravitational radiation 236–7, 239, 242, 248–9, 526
 gravitational redshift 221, 242, 279–80, 526
 gravitational waves 236, 237, 239–40, 242, 526
 gravitation, law of 66–70, 71–2, 213, 214
 gravitons 95, 352, 355, 376, 526
 gravity 94–5, 95–6, 116, 170, 191, 213, 237, 354, 526
 and acceleration 216–19
 as source of stars' energy 126
 quantum 254, 272, 355, 488–9, 507–8, 510, 511
 Grossman, Michael 178
 Guth, Alan 400, 474
 GUTs
 see grand unified theories
 hadron epoch 360, 377, 526
 hadrons 356–7, 359, 376, 526
 Hahn, Otto 91
 half-life 91, 526
 Halley, Edmund 59, 71, 407
 Halley's Comet 71
 harmonic law 48–9
 Harrison, Edward R. 408
 Harrison-Zel'dovich
 spectrum 447–8, 461, 526
 Harvard Smithsonian Observatory 451
 Hawking radiation 259–62, 272, 349, 479, 503, 526
 Hawking, Stephen 155, 256, 259
 Hawkins, I. 411
 HDM
 see hot dark matter
 heat 96–7, 116, 526
 caloric theory of 19–20
 heat death 371, 377, 526
 heavy hydrogen
 see deuterium
 Heisenberg uncertainty principle 259–60, 272, 491, 534
 Heisenberg, Werner 93
 heliocentric model 55, 526
 Aristarchus 34–5
 Copernican 39–42
 helium 342
 abundance 367–8
 creation 363–4
 isotope 365
 Helmholtz, Hermann 77, 126
 Henderson, T. 45
 Herman, Robert 339, 340, 363, 411
 Herschel, John 408
 Herschel, William 72, 281
 Hertzsprung-Russell diagram (HR diagram) 132, 134, 135
 Hewish, Anthony 141
 Higgs bosons 446, 473–4, 475, 476–7, 482–3, 526
Homo (genus) 78
 homogeneity 160–1, 179, 302, 526
 Hooke, Robert 58–9
 horizon 465, 484, 526
 horizon problem 464–8, 478, 483, 484, 526
 Horsehead Nebula 121
 hot dark matter (HDM) 399, 404, 445, 461, 526
 see also neutrinos
 hot dark matter (HDM) model 457–8, 526
 Hoyle, Fred 157, 338, 341
 HR diagram
 see Hertzsprung-Russell diagram
HST
 see Hubble Space Telescope
 Hubble constant 291, 293, 303–4, 308, 309, 320–1, 323, 380, 526
 and redshift-distance relationship 382–7
 Hubble Deep Field 386, 438

- Hubble, Edwin 287, 288,
289, 308,
339, 343
- Hubble expansion 293, 295,
309, 526
- Hubble flow 421, 422, 449,
527
- Hubble law 290–5, 303,
307, 308, 309, 527
and scale factor 320–2
- Hubble length 305, 310,
448, 465, 527
- Hubble Space Telescope*
(*HST*) 292, 382,
385, 397, 438, 459
- Hubble sphere 305, 310,
527
- Hubble time 303–4, 310,
527
and age of universe
324–5, 387–8
- Hulse, Russell 239
- Humason, Milton 289, 343
- Hutton, James 74
- Huygens, Christian 102
- Hydra constellation 399
- hydrostatic equilibrium
126, 129, 131, 147,
527
- hyperbolic geometry 227–8,
229, 242, 527
- hypothesis 18, 20, 22, 527
criterion 18–20, 22, 23
- ideal gas 147, 527
- ideal gas law 127–8, 147,
527
- IMF
see initial mass function
- impetus theory 32–3
- induction 17, 18
- inert gases 89
- inertia 30, 53, 55, 167, 179,
527
- inertia, law of 60–1
- inertial forces 169, 179,
214–15, 216, 527
- inertial motion 167, 179,
223, 527
- inertial observer 171–2,
179, 527
- inertial reference frame
168, 170–1, 173, 179,
184–5, 200, 203–4,
527
- inflation 356, 376, 400–1,
472, 475–7, 483–4,
527
and initial conditions of
universe 478–80
chaotic 480–1, 484
testing 482
- inflatons 477, 484, 527
- inheritance 74–5
- initial mass function (IMF)
123, 527
- interference 99–100, 116,
527
- interference fringes 100,
102, 493, 527
- interference pattern 492
- interferometer 175, 180,
527
- interstellar distances
measurement 284–7
- interstellar medium 120,
147, 527
- interstellar space 120
- invariance 172, 179, 527
- inverse square law 66
- ionic bond 89
- ion 89, 527
- ionized gases 89, 97
- irregular clusters 436, 527
- irregular galaxies 112, 436,
527
- island-universe hypothesis
282–3, 284
- isotopes 90, 115, 527
- isotropy 159–61, 179, 527
- James Webb Space*
Telescope 459
- Joule, Sir James 20, 96
- Jupiter 52, 58, 69, 72
- jupiters 396
- Kant, Immanuel 120
- Kelvin, Lord 77, 126
- Kepler, Johannes 47,
49–50, 54, 407
- Kepler's laws of planetary
motion 48–9, 55,
69–70
- Kerr black holes
see rotating black holes
- Kerr metric 257, 272, 528
- Kerr, Roy 257
- kinematical methods
393–4, 404, 528
- kinetic energy 96, 116, 194,
316, 502, 528
- Kirchoff, Gustav 107
- Lagrange point 138
- Lamarck, Jean-Baptiste 74,
75
- lambda density parameter
328, 334, 528
- Laplace, Pierre Simon de
121
- Large Magellanic Cloud
113, 289, 435, 441,
442, 444
- Larmor, Joseph 184
- Las Campanas Observatory
(Chile) 451
- Laser Interferometer
Gravitational Wave
Observatory (LIGO)
240
- Lavoisier, Antoine 85
- law 20, 528
- law of inertia
see inertia, law of
- Leavitt, Henrietta 287
- Lemaître, Georges 331,
339, 342
- Lemaître model 331–2, 334,
339, 528
- length contraction 184,
189–90, 191, 206,
209, 528
gravitational 221, 248
- Lense, J. 257
- Lense-Thirring effect 257
- lepton epoch 360–2, 377,
528
- leptons 356, 360–1, 376,
528
- Leverrier, Urbain 72
- Liebniz, Gottfried 223
- life (biology) 78–9
- light 102–3, 106–8, 277,
351, 489
bending of 256
extinction 109–10, 287
speed of 103–4, 181–2,
185, 191, 193,
207–8
- thought experiment
185–7
- lightcones 201, 202, 209,
528
- around black hole 251–2

- in the Robertson-Walker metric 466–7
- lightlike 200, 208, 209, 528
- lightyear 104, 116, 516, 528
- LIGO
see Laser Interferometer Gravitational Wave Observatory
- Linde, Andrei 477, 480
- linear momentum 62–3
- line radiation 107, 528
- Lippershey, Hans 50
- lithium
 isotope 365
- Lobachevsky, Nicholai 226
- Local Group (galaxy cluster) 113, 115, 293, 398, 420, 421, 422, 435–6, 459, 528
- Local Supercluster 436
- long range forces 191, 293, 528
- lookback time 307–8, 310, 528
- Lorentz contraction
see length contraction
- Lorentz-Fitzgerald contraction
see length contraction
- Lorentz, Hendrik 184
- Lorentz transformation 184, 185, 189, 190–2, 198, 209, 528
- Loys de Chesaux, Jean-Phillipe 408
- luminiferous ether
see ether
- luminosity 109, 115, 116, 528
 stellar 131–2
- luminosity distance 109–10, 286–7, 384, 309, 528
- Lyell, George 75
- M3 (globular cluster) 135
- M31
see Andromeda Galaxy
- M33 (spiral galaxy) 436
- M87 (elliptical galaxy) 112, 268, 270, 280, 436
- M100 (spiral nebula) 282, 293
- Mach, Ernst 183, 223
- MACHOs 441, 460, 461, 528
- Mach's principle 223–4, 232, 241, 242, 504, 528
- magnetic monopoles 470–1, 528
- main sequence 132, 133, 147, 528
- many-worlds interpretation 505–6, 511, 528
- Marduk (god) 8
- Mars
 distance to 44–5
 motion of 47–8
- mass 62, 68–9, 80, 179, 529
 inertial 167
 stellar 110
 unit 516
 universe 392–3
- mass density parameter 380–1
- mass-energy 195–6
- Massive Compact Halo Objects
see MACHOs
- mass-to-light ratio 394, 398, 404, 529
 for elliptical galaxies 395–6
 for spiral galaxies 395
- matter 85–6
 and energy 348–50
 conservation of 97
- matter density parameter 322–3, 334, 529
- matter era 345, 376, 529
- Maxwell, James Clerk 174
- Maxwell's equations 174, 182
- measurement problem 495, 511, 529
- measurements 165
see also cosmological parameters 380–2, 402–4
- mechanics 54, 55, 60, 529
 Aristotelian 29–30, 32–3, 52–3, 61
 Newtonian 18, 19, 60, 70–3, 80, 164, 178
see also Newton's laws of motion; quantum mechanics
- Meitner, Lise 91
- Mendeleev, Dmitry 86
- Mendel, Gregor 74–5, 76, 147
- mesons 357
- Messier 31 (M31)
see Andromeda Galaxy
- Messier, Charles 280
- Messier objects 280
- metal 134, 529
- metric coefficients 229, 242, 529
- metric equation 229–30, 242, 529
- Meyer, D. M. 411
- Meyer, Lothar 86
- Michelson, Albert 175
- Michelson-Morley experiment 176, 182–3, 184
- Michelson-Morley interferometer 175–6, 240
- microlensing 441–2, 529
- Milky Way Galaxy 51, 54, 78, 112, 113, 133, 162, 266, 280, 289, 308, 435, 529
 age 78
 motion 420–1
 size 281–2, 283–4
- millisecond pulsars 145–6
- Mills Methods 17
- Milne, E. A. 337
- Milne's model 337–8
- Minkowskian space-time 197, 200, 204, 209, 233, 234–5, 529
- Minkowski, Hermann 197
- missing mass
see dark matter
- Mizar 124
- Mizar B 124
- MKS unit of force
see newton
- models
see also scientific models
- molecular clouds 121–2
- moment of inertia 142–3
- monopoles 470
see also magnetic monopoles
- Montezuma II, Aztec emperor 10–11
- Morley, Edward W. 175
- muon neutrinos 443

- muons 196
 myth 6, 22, 529
 mythological cosmology 5–6, 7
 Babylonian 8
 Chinese 9–10
 Native American 10–11
 Tanzanian 8
- natural selection, theory of 75–6
 N-body algorithm 456
 nebulae 120–1, 147, 280–1, 309, 529
 spiral 281–3, 287, 288, 289, 308
- Neptune
 discovery 72
 neutralinos 446
 neutrino oscillation 443–4
 neutrinos 92, 115, 129, 361–2, 442–3, 444–5, 448, 461, 529
 types 443
 neutron degeneracy 140–1, 148, 529
 neutron degeneracy pressure 140, 529
 neutrons 90
 neutron stars 140–1, 142, 143–4, 148, 529
 in binary systems 144–5, 146
 newton 69
 Newtonian dark star 245, 249
 Newtonian equivalence principle 217–19, 242, 529, 535
 Newtonian mechanics 18, 19, 60, 70–3, 80, 164, 178
 Newtonian universe 314–17
 Newton, Isaac 57–60, 79, 102, 219, 407, 489
 Newton's laws of motion 18, 19, 60, 70–1
 first law 60–1, 80, 529
 second law 61–4, 80, 529
 third law 64–5, 80, 529
 Newton's rings 102
 NGC 3370 (spiral galaxy) 111
 NGC 4261 (galaxy) 269
- noble gases
 see inert gases
- Noddack, Walter 91
- no-hair theorem 257, 271, 272, 529
- novae 138–9, 148, 282, 309, 529
- Novikov, I. D. 411
- nuclear fission 91–2, 525
 nuclear forces
 see strong nuclear force, weak nuclear force
- nuclear fusion 92, 196, 525
 and stellar temperature 129–30
- nuclear physics 77, 90
 nuclear reactions 77, 90–2, 115, 126–7, 530
- nuclear weapons 195
- nucleons 90, 115, 530
- nucleosynthesis 339–41, 375, 376, 399, 530
 stellar 341–2
- nucleosynthesis epoch 362–3, 377, 530
- nucleus 88, 89, 90–92, 530
- Nu Wa (goddess) 9
- null *see* lightlike
- Occam's Razor 20, 22, 156, 530
- Olbers, Heinrich 408
- Olbers' paradox 408, 430, 431, 530
- omega *see* density parameter
- open universe 325, 334, 530
- Oppenheimer, J. Robert 141
- optics 58
- orbital velocities
 mass and 394–5
- Origin of Species, The* (Darwin) 76
- Orion Nebula 121, 280
- pair production 349, 376, 530
 threshold temperature 349–50
- Pan Gu (mythological character) 9
- parallax 35, 43, 55, 284–6, 309, 530
- diurnal 44
- stellar 45–6
- parsec 285, 309, 516, 530
- Parsons, William
 see Rosse, Lord
- particle horizon 465, 466–7, 484, 530
- particle physics 93–5, 114, 351, 353–4
- past 201, 202, 209, 530
- Pauli, Wolfgang 92, 94
- peculiar velocity 293, 309, 420, 530
 distortions 452–3
- Peebles, P. James E. 410, 411
- Penrose, Roger 256, 503
- Penzias, Arno 344, 409, 411, 416
- perfect cosmological principle 163, 179, 530
- Periodic Table of elements 86, 87, 89, 530
- perpetual-motion machines 97–8
- phase transition 354
- Philosophiae Naturalis Principia Mathematica* (Newton) 59, 60, 79
- photoelectric effect 177, 489
- photon damping 427, 431, 448, 461, 530
- photon diffusion 427, 448
- photon orbit 255
- photons 88, 95, 106, 115, 489, 530
 energy of redshifted 414–15
 in expanding universe 346–7
 traveling 412–14
- photon sphere 255–6, 272, 530
- physical constants 519
- pions 95
- Planck energy 472–3
- Planck epoch 355, 376, 471, 488, 503, 530
- Planck length 355, 530
- Planck, Max 93, 106, 177, 489
- Planck* satellite 430, 482
- Planck time 355, 530

- Planck's constant 106, 489, 530
 planetary motion
 Kepler's laws of 47–9, 69–70
 planetary orbits 238
 planetary systems
 formation of 124–5
 see also solar system
 planetesimals 122
 planetoids 122–3
 plasma 368, 530
 Plato 27, 29
 Poincaré, Henri 184
 Popper, Karl 18–19
 Population I (Pop I) 133–4, 147, 530
 Population II (Pop II) 134, 147
 Population III (Pop III) 134, 147, 530
 positrons 94, 488, 530
 potential 474, 482–3, 484, 530
 and scalar field 474–5
 potential energy 96, 116, 531
 predictive power 20, 23, 531
 Priestly, Joseph 85
 primordial elements 130–1, 147, 531
 primordial stars 134
 principle of causality 202, 209, 531
 principle of reciprocity 188–9, 190, 209, 531
Principles of Geology (Lyell) 75
 proper length 189, 208, 209, 531
 proper time 188, 199, 209, 499, 531
 proton-proton process 130
 protons 88, 90, 129, 531
 life expectancy 357–8
 protostars 122, 123
 Ptolemaic cosmology 35–7, 40, 78
 Ptolemy 35
 pulsars 141–2, 144, 148, 531
 binary 239
 millisecond 145–6
 Pythagorean theorem 166, 198
 QCD
 see quantum chromodynamics
 QSO
 see quasars
 quantum 89, 531
 quantum chromodynamics (QCD) 358
 quantum cosmology 496–8, 511, 531
 quantum fluctuations 472, 484, 531
 quantum gravity 254, 272, 355, 488–9, 507–8, 510, 511, 531
 quantum mechanics 93–4, 106, 114, 351–2, 376, 487, 490–2, 496, 506–7, 508–9, 531
 quantum state 94, 490, 511, 531
 quantum tunneling 497
 quark epoch 358–60, 377, 531
 quarks 95, 356, 377, 531
 quasars 145, 267, 268, 270, 273, 343–4, 437–8, 531
 mirrored 332
 quasi-stellar objects (QSO)
 see quasars
 Quetzalcoatl (serpent god) 10
 quintessence 402, 404, 531
 radiation 104, 115, 531
 line 107
 thermal 104–5
 radiation era 345, 347, 368–70, 375, 376, 423–4, 531
 radiation pressure 345–6
 radioactive dating 77, 78, 80, 91, 531
 radioactivity 90, 531
 radio astronomy 343
 radio galaxies 268
 radio jets 268–9
 radio lobes 268
 radio waves 102
 reception distance 383–4, 404, 531
 reciprocity, principle of
 see principle of reciprocity
 recombination 370, 377, 412, 416, 431, 531
 red dwarfs 123, 396, 531
 red giant 136, 147, 148, 531
 redshift 101–2, 115, 116, 194, 267, 278, 279, 309, 384–5, 531
 galaxy 451–2
 redshift-distance relationship 289–90, 383–4, 404, 531
 redshift-temperature relationship 413–14
 regular cluster 436, 527, 532
 reheating 356
 reionization 429, 431, 437, 460, 532
 relativity 172, 179–80, 487, 532
 postulates 184–5
 see also general theory of relativity; special theory of relativity
 relativity of uniform motion 53, 60
 relativity principle 184, 208, 209, 532
 relevance (of scientific theory) 18, 22, 532
 relic problem 470–1, 478, 483, 484, 532
 rest energy 195–6, 209, 532
 rest mass 194, 532
 retrograde motion 36, 55, 532
Revolutionibus orbium coelestium, De (On the Revolution of Heavenly Spheres) (*Copernicus*) 39
 Riemann, Georg F. B. 235
 Riemannian geometry 178, 235, 242, 532
Rig Veda 7
 Robertson, H. P. 339
 Robertson-Walker metric 301–2, 305, 309, 310, 321, 464, 532
 Roche lobes 138
 Roll, P. G. 410
 Rosse, Lord 281

- rotating black holes 256–9, 270–1
 Roth, K. C. 411
 Rutherford, Ernest 87–8
 Ryle, Martin 343
- Sachs-Wolfe effect 424, 431, 532
 Sagittarius A* (Sgr A*) 266
 Sandage, Allan 343
Scala Naturae (Aristotle) 75
 scalar field 474–5
 scale factor 299, 310, 499, 532
 and Hubble law 320–2
 Schmidt, Maarten 267, 343
 Schrödinger equation 490–1, 510, 511, 532
 Schrödinger, Erwin 93
 Schrödinger's cat 494–5, 506, 510
 Schwarzschild, Karl 246, 271
 Schwarzschild metric 246–8
 Schwarzschild radius 247, 248–9, 250, 271, 272, 532
 scientific law 20, 22, 528
 scientific methods 17–18, 22, 26, 532
 scientific models 21, 22, 529
 scientific notation 513–14
 scientific theory 20–1, 22, 533
 Scott, David 171
 second law of
 thermodynamics 97–98, 114, 372–373, 376, 377, 500, 532
 seed perturbations 423, 431, 469, 479, 532
 selective memory 13
 Seyfert galaxies 265, 268
 Sgr A*
 see Sagittarius A*
 Shane, C. D. 451
 Shapley, Harlow 283–4, 288
 Shklovsky, I. S. 411
 simplicity (of scientific theory) 20, 22, 532
 simultaneity 185, 201, 209, 532
 simultaneous events 201
 singularity 250, 272, 374, 532
 of black hole 253–4, 488
 Slipher, Vesto 282, 289, 298
 Sloan Digital Sky Survey 451, 452, 455
 Small Magellanic Cloud 289, 435, 441
 Smith, William 74
 SN1987A (supernova) 444
SNAP
 see SuperNova Acceleration Probe
 solar luminosity 109, 394, 532
 solar mass 110, 394, 532
 solar system 110–11
 see also planetary systems
 sound horizon 425
 sound waves 99, 122, 181, 423–4, 425–6
 source term 504
 space
 physical nature 154–5
 spacelike 199, 209, 532
 space-time 197, 208, 209, 499, 532
 geometry of 232
 space-time diagram 197–8, 204, 209, 532
 space-time interval 198–200, 209, 246, 532
 special-relativistic quantum mechanics 488
 special theory of relativity 19, 177–8, 191–2, 208–9, 213, 215, 241
 clock paradox 205–8
 spectroscopy 278
 spectrum 99, 108, 116, 277–8, 532
 absorption 107
 electromagnetic 102–3, 277
 emission 107
 speed 179, 532
 of light 103–4, 181–2, 185–7, 191, 193, 207–8, 532
 spherical geometry 226–7, 242, 533
 spiral galaxies 111–12, 289, 533
 formation 439
 mass-to-light ratio 395
 spiral nebulae 281–3, 287, 288, 289, 308
 spontaneous symmetry breaking 353–4, 355, 376, 474, 533
 SS433 (star) 145
 stability gap 341
 standard candle 287, 309, 384, 385, 404, 533
 standard Friedmann models
 see standard models
 standard models 318–19, 322–6, 334, 342–3, 375, 379, 533
 starquake 141
Starry messenger, The (Galileo) 52
 stars 119–20, 126, 146–7, 533
 age 133–5
 birth 120–5, 429, 437
 blackbody radiation 106
 categories 133–4
 end of star's life 135–6
 energy generation 126–7
 luminosity 131–2
 mass 110, 131
 structure 128–9
 static surface 258, 272, 533
 steady state model 329–31, 338–9, 334, 341, 343, 374, 533
 Steinhardt, Paul 477
 stellar nucleosynthesis 341–2
 stellar parallax 45–6
 stellar remnants 397–8
 Steno, Nicolaus 74
 Strassmann, Fritz 91
 stress-energy tensor 236
 string theory 508, 511, 533
 strong equivalence principle
 see Einstein equivalence principle
 strong interactions 94, 95, 115–16, 529, 533
 strong nuclear force
 see strong interactions
 structure formation 369, 377, 447–8, 459–60, 478–9, 533

- and dark matter 448–9
 bottom-up 437, 460
 top-down 436–7, 460, 534
 structure problem 469–70,
 483, 484
 Struve, F. G. W. 45
 Sun 133
 age 77, 78, 80
 bending of starlight
 237–8
 distance from Earth 34
 end of Sun's life 135–6
 motion 422
 spectrum 107
 see also solar system
 sunspots 51
 superclusters 113, 115, 116,
 436, 460,
 533
 Superconducting
 Supercollider 351
 Super Kamiokande detector
 445
SuperNova Acceleration Probe (SNAP) 387
 supernovae 119, 139–40,
 148, 282, 309, 386–7,
 533
 Type Ia 140, 292, 385–6,
 393
 types 140
 superposition of states 492,
 494, 511,
 533
 supersonic aircraft 207
 boost factor 190
 surface of last scattering
 370, 416, 431, 464,
 533
 symmetries 353, 376, 533
- Tanaka, Y. 265
 Taylor, Joseph 239
 tau neutrinos 443
 telescope 50–1
 reflecting 57
 temperature 97, 116,
 499–500, 533
 black holes 261–2
 cosmic background
 radiation 418–19,
 421, 424–5
 stellar 129–30, 131–2
 threshold 349–50
 unit 516
- temperature-redshift
 relationship 413–14
 testability (of scientific
 theory) 18,
 22, 533
 Teucciztan (god) 10
 Tezcatlipoca (god) 10
 theory
 see scientific theory
 thermal equilibrium 344,
 376, 412, 533
 thermal radiation 104–5,
 533
 thermodynamics 16, 534
 and entropy 500–1
 laws 97–8, 114, 372, 377,
 532
 Thirring, H. 257
 Thomson, J. J. 86
 Thomson, William
 see Kelvin, Lord
 Thorne, Kip 263, 504
 thought experiments 52,
 55, 185, 534
 Tiamat (goddess) 8
 tidal forces 233–4, 242,
 503, 534
 around black holes 250–1
 time 498–500
 measurement 191
 physical nature 154–5
 proper 188, 199, 531
 unit 516
 see also space-time
 time dilation 186–8, 191,
 206, 209, 534
 gravitational 221–2,
 247–8
 timelike 199, 208, 209, 504,
 534
 time travel 504–6
 Tolman, Richard 339
 top-down structure
 formation 436–7,
 460, 534
 torsion balance 219
 transformation theory
 (biology) 74–5
 transuranic elements 91–2
 triple-alpha process 136,
 157, 341
 true vacuum 476, 484, 534
 Tully-Fisher relation
 292–3, 309, 534
 Tully, R. Brent 292
- tunneling 129
 quantum 497
 turnoff mass 135, 147–8,
 534
 twin paradox 203–5, 223
 Two-Micron All Sky Survey
 (2MASS) 455
 Tycho Brahe 42, 43–5, 46,
 47
- uncertainty principle
 see Heisenberg
 uncertainty
 principle
 Uhuru (X-ray satellite) 264
 ultraviolet catastrophe 106
 ultraviolet rays 102, 103
 unified epoch 356–7, 376,
 534
 uniformitarianism 74, 75
 uniform motion 60–1, 80,
 534
 universal gravitation,
 66–70, 71–2, 80, 534
 universe 3–4, 5, 22, 25, 153,
 158, 308, 534
 age 78, 304, 324–5, 387–9
 cyclic 374–5, 503
 end of 370–2
 entropy 502–3
 evolution 155–6, 343–4,
 375–6
 expanding 297–304,
 337–8, 408–9
 geometry 389–92, 403
 homogeneous 160–1,
 301–2
 initial conditions 156–7,
 478–80
 isotropic 159–61, 302–3,
 419
 mass 392–3
 models
 see cosmological
 models
 multiple universes 158
 Newtonian 314–17
 origin 508–9
 parameters 380–2, 402–3,
 404, 530
 physical 153–5, 178–9
 simulation
 see cosmological
 simulation
 size 305

- Uranus
discovery 72
- vacuum breakdown 260
vacuum energy 401, 404,
472, 483–4, 534
vacuum polarization 260
van Maanen, Adriaan
282–3
vectors 63, 80, 534
velocity 61, 63, 69, 166,
167, 179, 534
escape 245, 272, 316, 525
orbital 394–5
peculiar 293, 309, 420,
452–3, 530
relativistic addition
192–3
unit 516
- Venus
phases 51, 52, 54
- Virgo Cluster 113, 292,
293, 382–3, 398–9,
420, 436, 534
- virial theorem 395, 404,
534
- virtual particles 260, 272,
534
- voids 451, 461, 534
- Volkoff, George 141
- Wallace, Alfred Russell 76
- water waves 99
- wave amplitude 99, 425,
426, 427, 535
- wave frequency 103, 535
wavefunction 490–2, 511,
535
collapse 492–3, 496, 511,
522
- wavefunction of the
universe 497, 511,
535
- wave interference
see interference
- wave length 99, 101, 103,
115, 248, 278,
305–306, 346, 413,
535
- wave-particle duality
351–3, 489–90
- wave reflection 99
- wave refraction 99
- waves 99–102, 114–15,
351–2, 534
- weak equivalence principle
see Newtonian
equivalence
principle
- weak interactions 94, 95,
116, 529, 535
- weakly interacting massive
particles
see WIMPs
- weak nuclear force
see weak interactions
- weight 68–9, 80, 535
- weightlessness 214, 215–17
- Wheeler-De Witt equation
496–7, 498
- Wheeler, John 496
- white dwarfs 136–9, 146,
148, 239, 535
- white holes 262
- white noise 425
- Wien displacement law 105
- Wilkinson, D. T. 410
- Wilkinson Microwave
Anisotropy Probe*
(*WMAP*) 382, 390,
427–30, 437, 449, 482
- Wilson, Robert 344, 409,
411, 416
- WIMPs 446, 448, 461, 535
- Wirtanen, C. A. 451
- WMAP*
see *Wilkinson Microwave
Anisotropy Probe*
- Woolf, N. J. 411
- work 96, 114, 116
- worldline 197, 199, 209,
251, 499, 504, 535
- wormholes 262–3, 504
- Wren, Christopher 59
- Wright, Thomas 280
- Xolotl (god) 10
- X-ray bursters 145
- X-ray surveys 456
- ylem 340
- Young, Thomas 102
- Zel'dovich, Ya. B. 411
- Zeta Ophiuchus (star) 411
- Zwicky, Fritz 141