The background of the book cover features a complex, abstract pattern of wavy lines composed of small dots, resembling a digital or quantum field. The colors transition from deep purple at the top to orange and red towards the bottom, creating a sense of depth and motion.

Delia Perlov
Alex Vilenkin

COSMOLOGY FOR THE CURIOUS



Springer

Cosmology for the Curious

Delia Perlov · Alex Vilenkin

Cosmology for the Curious



Delia Perlov
Tufts University
Medford, MA, USA

Alex Vilenkin
Tufts University
Medford, MA, USA

ISBN 978-3-319-57038-9 ISBN 978-3-319-57040-2 (eBook)
DOI 10.1007/978-3-319-57040-2

Library of Congress Control Number: 2017938144

© Springer International Publishing AG 2017

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

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

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

Printed on acid-free paper

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

To the memory of Allen Everett and Leonard Schwartz

Acknowledgements

We would like to express our sincere thanks to the Springer publishing team, and especially to Angela Lahee. Angela has been extremely helpful, accommodating and patient at each step of the way. We would like to thank the following people for reading some or all of the manuscript and offering useful feedback: Jose Blanco-Pillado, Peter Jackson, Jim Kernohan, Levon Pogosian, Michael Schneider and Brian Sinskie. A special thank you to Ken Olum for his extensive comments. Thanks also to Natalie Perlov for drawing several figures in the book, and to Gayle Grant and Caroline Merighi at Tufts University for their administrative help. DP: I wish to thank my husband Larry, my children Natalie, Alexa and Chloe, my mother Glenda, sister Heidi, and my late father Leonard for continued support and interest in this project. AV: It would have been hard to get to the end of this project without the support I had from my wife Inna. I thank her for her patience, advice, and for the wonderful cuisine that kept up my spirits.

Contents

Part I The Big Bang and the Observable Universe

1 A Historical Overview	3
1.1 The Big Cosmic Questions	3
1.2 Origins of Scientific Cosmology	4
1.3 Cosmology Today	7
2 Newton's Universe	13
2.1 Newton's Laws of Motion	13
2.2 Newtonian Gravity	16
2.3 Acceleration of Free Fall	19
2.4 Circular Motion and Planetary Orbits	20
2.5 Energy Conservation and Escape Velocity	22
2.6 Newtonian Cosmology	26
2.7 Olbers' Paradox	27
3 Special Relativity	31
3.1 The Principle of Relativity	31
3.2 The Speed of Light and Electromagnetism	35
3.3 Einstein's Postulates	39
3.4 Simultaneity	41
3.5 Time Dilation	42
3.6 Length Contraction	44
3.6.1 Speeding Muons	45
3.7 $E = mc^2$	46
3.8 From Space and Time to Spacetime	47
3.9 Causality in Spacetime	51

4	The Fabric of Space and Time	59
4.1	The Astonishing Hypothesis	60
4.2	The Geometry of Space	63
4.2.1	Euclidean Geometry	63
4.2.2	Non-Euclidean Geometry	66
4.3	Curved Space	67
4.3.1	The Curvature of Surfaces	67
4.3.2	The Curvature of Three-Dimensional Space	70
4.4	The General Theory of Relativity	72
4.5	Predictions and Tests of General Relativity	75
4.5.1	Light Deflection and Gravitational Lensing	75
4.5.2	Gravitational Time Dilation	77
4.5.3	Black Holes	77
4.5.4	Gravitational Waves	78
5	An Expanding Universe	83
5.1	Einstein's Static Universe	83
5.2	Problems with a Static Universe	86
5.3	Friedmann's Expanding Universe	89
6	Observational Cosmology	97
6.1	Fingerprints of the Elements	98
6.2	Measuring Velocities	99
6.3	Measuring Distances	101
6.4	The Birth of Extragalactic Astronomy	105
7	Hubble's Law and the Expanding Universe	109
7.1	An Expanding Universe	110
7.2	A Beginning of the Universe?	113
7.3	The Steady State Theory	114
7.4	The Scale Factor	115
7.5	Cosmological Redshift	116
7.6	The Age of the Universe	117
7.7	The Hubble Distance and the Cosmic Horizon	118
7.8	Not Everything is Expanding	120
8	The Fate of the Universe	125
8.1	The Critical Density	125
8.2	The Density Parameter	128

9 Dark Matter and Dark Energy	131
9.1 The Average Mass Density of the Universe and Dark Matter	131
9.2 Dark Energy	136
9.3 The Fate of the Universe—Again	140
10 The Quantum World	143
10.1 Quantum Discreteness	143
10.2 Quantum Indeterminism	145
10.3 The Wave Function	148
10.4 Many Worlds Interpretation	151
11 The Hot Big Bang	155
11.1 Following the Expansion Backwards in Time	155
11.2 Thermal Radiation	158
11.3 The Hot Big Bang Model	161
11.4 Discovering the Primeval Fireball	162
11.5 Images of the Baby Universe	165
11.6 CMB Today and at Earlier Epochs	168
11.7 The Three Cosmic Eras	170
12 Structure Formation	175
12.1 Cosmic Structure	175
12.2 Assembling Structure	179
12.3 Watching Cosmic Structures Evolve	180
12.4 Primordial Density Fluctuations	182
12.5 Supermassive Black Holes and Active Galaxies	183
13 Element Abundances	187
13.1 Why Alchemists Did Not Succeed	187
13.2 Big Bang Nucleosynthesis	189
13.3 Stellar Nucleosynthesis	193
13.4 Planetary System Formation	194
13.5 Life in the Universe	196
14 The Very Early Universe	201
14.1 Particle Physics and the Big Bang	201
14.2 The Standard Model of Particle Physics	205
14.2.1 The Particles	206
14.2.2 The Forces	206
14.3 Symmetry Breaking	208
14.4 The Early Universe Timeline	211

14.5	Physics Beyond the Standard Model	213
14.5.1	Unifying the Fundamental Forces	213
14.6	Vacuum Defects	215
14.6.1	Domain Walls	216
14.6.2	Cosmic Strings	217
14.6.3	Magnetic Monopoles	220
14.7	Baryogenesis	220

Part II Beyond the Big Bang

15	Problems with the Big Bang	227
15.1	The Flatness Problem: Why is the Geometry of the Universe Flat?	227
15.2	The Horizon Problem: Why is the Universe so Homogeneous?	229
15.3	The Structure Problem: What is the Origin of Small Density Fluctuations?	232
15.4	The Monopole Problem: Where Are They?	232
16	The Theory of Cosmic Inflation	235
16.1	Solving the Flatness and Horizon Problems	235
16.2	Cosmic Inflation	236
16.2.1	The False Vacuum	236
16.2.2	Exponential Expansion	238
16.3	Solving the Problems of the Big Bang	240
16.3.1	The Flatness Problem	240
16.3.2	The Horizon Problem	241
16.3.3	The Structure Formation Problem	242
16.3.4	The Monopole Problem	242
16.3.5	The Expansion and High Temperature of the Universe	242
16.4	Vacuum Decay	243
16.4.1	Boiling of the Vacuum	243
16.4.2	Graceful Exit Problem	244
16.4.3	Slow Roll Inflation	245
16.5	Origin of Small Density Fluctuations	247
16.6	More About Inflation	249
16.6.1	Communication in the Inflating Universe	249
16.6.2	Energy Conservation	250

17 Testing Inflation: Predictions and Observations	255
17.1 Flatness	255
17.2 Density Fluctuations	256
17.3 Gravitational Waves	260
17.4 Open Questions	264
18 Eternal Inflation	269
18.1 Volume Growth and Decay	269
18.2 Random Walk of the Inflaton Field	271
18.3 Eternal Inflation via Bubble Nucleation	274
18.4 Bubble Spacetimes	275
18.5 Cosmic Clones	279
18.6 The Multiverse	281
18.7 Testing the Multiverse	284
18.7.1 Bubble Collisions	284
18.7.2 Black Holes from the Multiverse	285
19 String Theory and the Multiverse	291
19.1 What Is String Theory?	292
19.2 Extra Dimensions	294
19.3 The Energy Landscape	295
19.4 String Theory Multiverse	296
19.5 The Fate of Our Universe Revisited	297
20 Anthropic Selection	301
20.1 The Fine Tuning of the Constants of Nature	302
20.1.1 Neutron Mass	302
20.1.2 Strength of the Weak Interaction	303
20.1.3 Strength of Gravity	303
20.1.4 The Magnitude of Density Perturbations	303
20.2 The Cosmological Constant Problem	304
20.2.1 The Dynamic Quantum Vacuum	304
20.2.2 Fine-Tuned for Life?	305
20.3 The Anthropic Principle	307
20.4 Pros and Cons of Anthropic Explanations	309
21 The Principle of Mediocrity	313
21.1 The Bell Curve	313
21.2 The Principle of Mediocrity	314
21.3 Obtaining the Distribution by Counting Observers	315

21.4	Predicting the Cosmological Constant	316
21.4.1	Rough Estimate	317
21.4.2	The Distribution	317
21.5	The Measure Problem	319
21.6	The Doomsday Argument and the Future of Our Civilization	321
21.6.1	Large and Small Civilizations	322
21.6.2	Beating the Odds	323
22	Did the Universe Have a Beginning?	327
22.1	A Universe that Always Existed?	327
22.2	The BGV Theorem	329
22.2.1	Where Does This Leave Us?	330
22.2.2	A Proof of God?	331
23	Creation of Universes from Nothing	333
23.1	The Universe as a Quantum Fluctuation	333
23.2	Quantum Tunneling from “Nothing”	336
23.2.1	Euclidean Time	337
23.3	The Multiverse of Quantum Cosmology	338
23.4	The Meaning of “Nothing”	339
24	The Big Picture	343
24.1	The Observable Universe	343
24.1.1	What Do We Know?	343
24.1.2	Cosmic Inflation	344
24.2	The Multiverse	345
24.2.1	Bubble Universes	345
24.2.2	Other Disconnected Spacetimes	346
24.2.3	Levels of the Multiverse	346
24.2.4	The Mathematical Multiverse and Ockham’s Razor	347
24.3	Answers to the “Big Questions”	350
24.4	Our Place in the Universe	351
Appendix A		353
Further Reading		361
Index		365

Part I

The Big Bang and the Observable Universe

1

A Historical Overview

1.1 The Big Cosmic Questions

Cosmology is the study of the origin, nature and evolution of our universe. Its practitioners strive to describe cosmic history in quantitative detail, using the language of modern physics and abstract mathematics. Yet, at its core, our cosmological knowledge is the answer to a few fundamental questions. Have you ever drifted off deep into thought, wondering: Is the universe finite or infinite? Has it existed forever? If not, when and how did it come into being? Will it ever end? How do we humans fit into the grand scheme of things? All ancient and modern cultures have developed creation stories where at least some of these questions have been addressed.

In one of the Chinese creation myths, the universe begins as a black egg containing a sleeping giant, named Pan Gu. He slept for 18,000 years and grew while he slept. Then he woke up and cracked the egg open with an ax. The light part of the egg floated up to form the sky, while the heavy part stayed down and formed the Earth. Pan Gu remained in the middle and continued to grow, pushing the sky and the Earth further apart. When Pan Gu died, his breath became the wind, his eyes the Sun and the Moon, his sweat turned into rain, and the fleas in his hair transmuted into humans.

The prospect of being a descendant of fleas may not be fully satisfying, but perhaps an even more objectionable aspect of this story is that it does not address the obvious question: “Where did the black egg come from in the first place?” Similar types of questions also arise in the context of scientific cosmology. Even if we claim to know what happened at the beginning of the universe, you can always ask: And what happened before that?

There is also a limit to how far we can see in space, so how can we know what lies beyond?

For a long time it seemed as though we would never know the answers to the “big” cosmic questions. Thus, cosmologists focused mostly on the part of the universe that could be directly observed, leaving it to philosophers and theologians to argue about the great mysteries. We shall see, however, that due to remarkable developments in cosmology over the last few decades, we now have answers, that we have reason to believe, to at least some of the big questions.

1.2 Origins of Scientific Cosmology

The idea that the universe can be rationally understood is at the foundation of all scientific knowledge. This concept is now commonplace, but in Ancient Greece more than 20 centuries ago it was a daring hypothesis. The Greek philosopher Thales (6th century BC) suggested that all of Nature’s variety could be understood from a few basic principles, without the intervention of gods. He believed that the primary element of matter was water. Two centuries later, Democritus advocated that all matter was made up of tiny, eternal, indivisible particles, called atoms, which moved and collided with one another in empty space. He stated: “Nothing exists except atoms and empty space.” This line of thought was further developed by Epicurus (3rd century BC), who argued that complex order, including living organisms, evolved in a natural way, by random collisions and rearrangements of atoms, without any purpose or intelligent design. Epicurus asserted that atoms occasionally experience small random “swerves” from their rectilinear motion. He believed that these deviations from strict determinism were necessary to explain the existence of free will. Epicurus taught that the universe is infinite and that our Earth is just one of countless worlds that constantly form and decay in an infinite space (Fig. 1.1).

Another important direction of thought originated with Pythagoras (6th century BC), who believed that mathematical relations were at the heart of all physical phenomena. Pythagoras was the first to call the heavens *cosmos*, which means *order*. He suggested that the Earth, the Sun, and other celestial bodies are perfect spheres and move in perfect circles around a central fire, which cannot be seen by human eyes. Think about how different this is from the random aggregates of atoms envisioned by Epicurus!

In the 4th century BC, Plato and then Aristotle proposed more elaborate versions of this picture, placing the Earth at the center of the universe, with

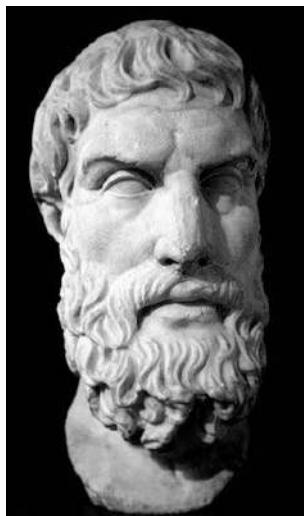


Fig. 1.1 Epicurus (341–270 BC) taught philosophy in the garden of his house in Athens, where he regularly met with a small group of followers over a simple meal. The group included women and one of his slaves. Epicurus was a prolific writer, but almost all of his writings have vanished. Epicurean philosophy flourished in ancient Greece and Rome for several centuries, but was banished in the Christian world, because of its uncompromising materialism. Its most complete exposition came to us in a magnificent poem “On the Nature of Things”, written in the first century AD by the Roman poet Lucretius. The poem was lost for more than a thousand years and was rediscovered in a German monastery in 1417, just in time to influence the development of ideas during the Renaissance

the planets, the Sun and the stars attached to translucent spheres rotating about the center. This was a decidedly finite universe, where the stars were placed on the outermost sphere.

The Greeks made very accurate observations of the planets, and already in the 3rd century BC it had become evident that the simple model of concentric spheres could not adequately explain the observed motion of the planets. Further refinements of the model were getting more accurate, at the expense of becoming more complicated. First, the centers of the spheres were displaced by certain amounts from the Earth. Then came the idea of epicycles: each planet moves around a small circle, whose center rotates around a large circle, as shown in Fig. 1.2. Epicycles explained why planets seem to move backward and forward on the sky, and why they appear to be brighter during the periods of backward motion.

In some cases epicycles had to be added on top of other epicycles. All of these ideas were consolidated by Claudius Ptolemy in his book *Almagest* (The Great System), in the 2nd century AD. Ptolemy’s mathematical model

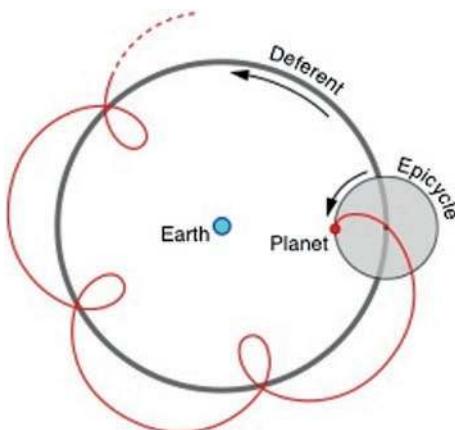


Fig. 1.2 The planet moves around a *small circle* (epicycle), whose center moves around a *large circle* (deferent) centered on Earth. The planet's resulting trajectory is shown here in red; most of the time the planet moves in the "forward" direction relative to the background stars, but for brief intervals, when the planet is close to the Earth, and hence is at its brightest, its direction of motion is reversed relative to the background stars. Credit Daniel V. Schroeder

of the universe endured for fourteen hundred years. It accounted for all known astronomical data and also made accurate predictions.

The dismantling of the Ptolemaic worldview began in the 16th century with the work of Nicolaus Copernicus. He wanted to restore the ideal of perfect circular motion by placing the Sun at the center of the universe, and allowing the Earth to move around it in a circular orbit (this idea actually goes back to Aristarchus in the 3rd century BC). As the Earth circles around the Sun, the planets appear to move backward and forward across the sky, removing the "need" for epicycles. Copernicus devoted his life to the computation of heliocentric orbits and published his work in the book *On the revolutions of celestial spheres*, which came out in 1543, shortly before his death.

Despite its tremendous impact, it was not immediately clear that the Copernican system was superior to that of Ptolemy. Copernicus discovered that the simple model of circular orbits did not fit the data well enough. Ultimately, he also had to introduce epicycles, and even then he could not match the accuracy of Ptolemy's *Almagest*. Despite these setbacks, Copernicus still deserves to be immortalized for his greatest achievement—removing the Earth from the center of the universe. It has been downhill for the Earth ever since then,¹ but more on that later.

¹In fact, removing the Earth from the center of the universe was not necessarily viewed as a demotion. In those days the further out you went from the center, the closer you got to the heavenly celestial realm.

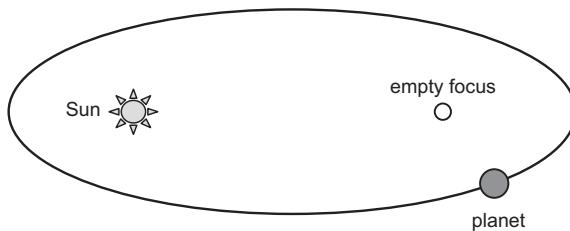


Fig. 1.3 Kepler discovered that planetary orbits are ellipses. (What is an ellipse? Consider two points, called the foci. An ellipse is the locus of points such that the sum of the distances to each focus is constant.) The Sun is located at one of the focal points of the ellipse, while the other focus is empty. For planets in the Solar System, the two foci of the ellipse are very close to one another, so the orbits are nearly circular. In this figure the ellipse is exaggerated

The next great astronomical breakthrough was made by Johannes Kepler in the early 1600s. After nearly three decades of studying the data amassed by his eccentric mentor Tycho Brahe, Kepler discovered that planets actually move along elliptical orbits. He realized the importance of his work, but was still very disappointed, because he believed that circles are more perfect than ellipses. Kepler had other mystical beliefs—in answer to the mystery of why each planet followed its particular orbit, he suggested that the planet grasped it with its mind! (Fig. 1.3).

Then along came Isaac Newton, who had very different ideas about how the laws of Nature operate. In his seminal book *Philosophiae naturalis principia mathematica* (1687), now known as the *Principia*, he showed how to derive the elliptical orbits of the planets from his three laws of motion and the law of universal gravitation. He postulated that the laws of Nature apply to all bodies, in all places and at all times. Newton's laws are mathematical equations that determine how physical bodies move from one moment to the next, describing a universe which functions like a giant clockwork mechanism. To set the clockwork up, one only needs to specify the initial conditions—the positions and velocities of all physical objects at some initial moment of time. Newton believed these were provided by God. We will return to Newton and his laws in some detail, but for now we jump ahead a few hundred years to outline what we know today.

1.3 Cosmology Today

Despite its ancient roots, scientific cosmology is a relatively young science. Most of what we know about the universe has been learned within the last 100 years. In broad-brush strokes, we have discovered that our Sun



Fig. 1.4 Andromeda Galaxy is one of our close neighbors at some 2.5 million light years away. It is about the same size as the Milky Way. Credit Robert Gendler

belongs to a huge disk-like conglomeration of about three hundred billion stars, known as the Milky Way galaxy. Not only is the Sun merely one out of hundreds of billions of stars in our galaxy, the Milky Way is itself only one out of hundreds of billions of galaxies that are scattered throughout the observable universe. Furthermore, Edwin Hubble showed (1929) that these distant galaxies are not just suspended at rest throughout space. Rather, they are rushing away from us, and each other, at very high speeds as the entire universe expands (Fig. 1.4).

If we extrapolate this expansion backwards in time, we realize that the universe was once much denser and much hotter. In fact, we believe that the universe as we know it originated some 14 billion years ago in a great explosion called the big bang. At that time, all of space was filled with an extremely hot, dense, and rapidly expanding “fireball”—a mixture of sub-atomic particles and radiation. As it expanded, the fireball cooled, along the way producing nuclei and atoms, stars and galaxies, you and us! In 1965, Arno Penzias and Robert Wilson discovered a faint remnant of the primordial fireball. They found that the entire universe is bathed in a sea of low-intensity microwaves,² known as the Cosmic Microwave Background radiation, or CMB.

²We are all familiar with x-rays, visible light and radio waves from our everyday lives. All of these are forms of electromagnetic radiation, which we will discuss later. Microwaves are a subset of radio waves.

Although the CMB had been predicted by theorists, Penzias and Wilson stumbled upon it serendipitously, providing the smoking gun proof for the big bang theory and earning themselves a Nobel prize in the process.

The big bang cosmology has its roots in Einstein's theory of gravity—the general theory of relativity (1915). Solutions of Einstein's equations describing an expanding universe were found by the Russian mathematician Alexander Friedmann (1922), and independently by the Belgian priest Georges Lemaitre (1927). The idea that the early universe was hot was introduced by the Russian expatriate George Gamow. Gamow wanted to explain the abundances of different chemical elements that we now observe in the universe. He argued that the hot primordial fireball was the furnace where the elements were forged by nuclear reactions. In 1948 Gamow and his colleagues Ralph Alpher and Robert Herman successfully calculated the abundances of hydrogen and helium produced during the big bang. They also tried to explain the abundances of heavier elements in the periodic table, but alas, here they were unsuccessful. It turns out that heavy elements are not synthesized during the big bang, but rather are produced in the interiors of stars. We will return to this part of our ancient history in more detail later. But suffice it to say, by the mid 1970s the major ingredients of the hot big bang picture were clearly outlined (Fig. 1.5).

Not so long ago, cosmology was not considered to be a reputable branch of science. There was very little data to test theoretical models. Two Nobel prize winning physicists, Lev Landau and Ernest Rutherford quipped, respectively, “Cosmologists are often in error, but never in doubt.” and “Don't let me catch anyone talking about the universe in my lab!” Attitudes changed dramatically in the 1980s and 90s, when an abundance of data emerged. Radio and optical astronomy flourished with computerized galaxy surveys and instruments like the Very Large Array (VLA) and the Cosmic Background Explorer (COBE) satellite. A detailed map of the distribution of galaxies in space has been compiled, showing remarkable large-scale structures of filaments, sheets and voids. The Hubble Space Telescope has captured images of galaxies so far away that it took much of the age of the universe for their light to reach us. By observing these distant galaxies we can see cosmic history unfolding. The turn of the century saw the launch of the Wilkinson Microwave Anisotropy Probe (WMAP) satellite, to further study the image of the early universe imprinted in the Cosmic Microwave Background radiation. All these developments (and others) ushered in an era of unprecedented precision cosmology, and we are fortunate to find ourselves living during this golden age! (Fig. 1.6).

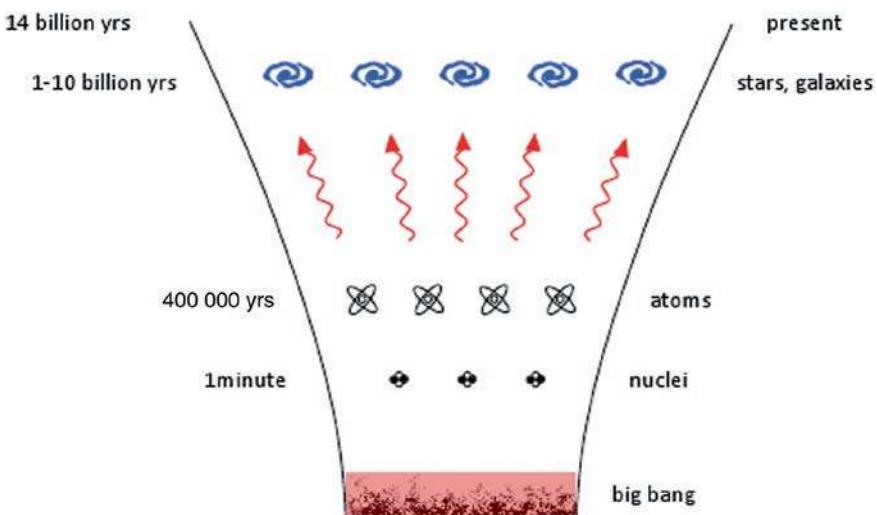


Fig. 1.5 Abridged history of the universe. Atomic nuclei were formed a few minutes after the big bang; four hundred thousand years later they combined with electrons to form atoms. At that point the universe became transparent to light, so we can see its image at that early era imprinted in the Cosmic Microwave Background radiation. Galaxies were pulled together by gravity over the course of several billion years, and we appeared on the scene in very recent cosmic time



Fig. 1.6 Very Large Array radio telescopes in New Mexico. Credit VLA, NRAO

While the hot big bang theory is supported by all observations, luckily for today's cosmologists, some intriguing questions still remain. These questions bring into play a combination of studies on the largest imaginable scales, and new theoretical insights from particle physics, on the smallest imaginable scales. From the microcosm to the macrocosm, our journey has begun...

Questions

What would be your answers (or best guesses) to the following questions:

1. Is the universe infinite or finite? If it is finite, does it have a boundary? If so, what lies beyond?
2. Did the universe have a beginning? If it did, was it an absolute beginning, or did the universe exist before that in some other form?
3. If the universe did have an absolute beginning, would that require a supernatural intervention?
4. Will the universe ever end? If so, will that be an absolute end, or will the universe be transformed into some other form?
5. What does the universe look like in far-away regions that we cannot observe? Is it similar to our cosmic neighborhood? Is our location in the universe in any way special?
6. Do you think the universe was designed to host intelligent life?
7. Do you think we are the only life in the universe?
8. Do you find it surprising that we are able to understand the universe? Do you find it surprising that mathematics is able to explain physical phenomena (like the elliptical orbits of planets)?
9. Do you think we have free will? If so, how can it coexist with deterministic laws of physics? Do the “swerves” of atoms posited by Epicurus give a satisfactory answer?

See if your answers change after you read this book!

2

Newton's Universe

In his monumental *Principia*, Newton formulated the general laws of motion and the law of universal gravitation. He then applied these laws to explain the motion of planets and comets, projectile trajectories, and the marine tides, among other things. In so doing, he showed how natural phenomena could be understood using a handful of physical laws, which hold just as well for the “heavenly Moon” as for the “Earthly apple” (Fig. 2.1).

2.1 Newton's Laws of Motion

Newton's first law states that a body that is at rest will stay at rest, and a body that is moving with a constant velocity will maintain that constant velocity, unless it is acted upon by a force.

What does this mean? Let's imagine we are at an ice rink and there is a hockey puck which has been carefully placed at rest on the ice. Now we stand and watch the puck. What happens? According to Newton, the puck will stay where it is unless someone comes by and gives it a push—that is, applies a force.¹

Now imagine we have given our little puck a push, so that it is sliding along the surface of the ice. We will assume that our ice rink has no friction. The puck will then continue to move at a constant speed in the same direction,

¹Even a motionless puck on frictionless ice is subject to forces. Gravity pulls the puck downwards, but the surface of the ice pushes back with equal and opposite force, so the total force on the puck is zero.

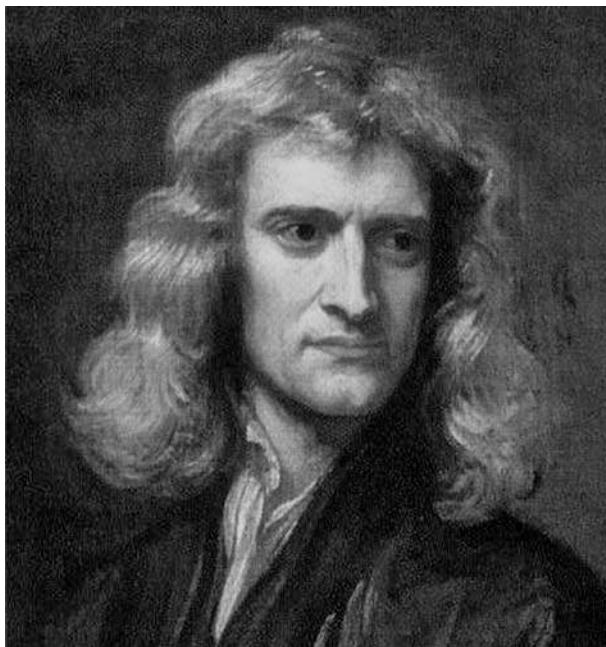


Fig. 2.1 Isaac Newton (1642–1726) made most of his major discoveries in 1665–1667, shortly after receiving his Bachelor's degree from the University of Cambridge. Although Newton earned financial support for further study, the University closed because of the plague, and he had to return to his family home in Lincolnshire for 18 months. It was during this time that he discovered his theory of colors, the law of gravitation, and calculus. In later years, apart from pursuing research in physics and mathematics, Newton devoted much effort to alchemy and to Scriptural studies. Credit Copy of a painting by Sir Godfrey Kneller (1689), painted by Barrington Bramley

unless it hits the wall of the rink, or bumps into someone or something along its way. These obstructions would provide a force that would alter the puck's uniform state of motion. If our imaginary frictionless ice rink were also infinite and devoid of other obstacles, the puck would coast along at the same velocity for eternity.

Newton's first law also goes by the name of *The Law of Inertia*.² A spaceship traveling with its engines turned off in interstellar space glides along

²The law of inertia was actually discovered by Galileo and was adopted by Newton as one of his laws of motion.

with a constant velocity, and provides yet another example of a body undergoing “inertial” motion.

Newton’s second law tells us that if a force is applied to a body, the body accelerates—meaning its velocity changes. The law can be stated mathematically as

$$\vec{a} = \vec{F}/m \quad (2.1)$$

where \vec{a} is the acceleration of the body, m is its mass, and \vec{F} is the applied force. The acceleration is defined as the rate at which the velocity changes. For example, if in one second the velocity changes by one meter per second, then the acceleration is one meter per second per second, or one meter per second squared (m/s^2). In general, if the velocity is in m/s , the acceleration is measured in m/s^2 .

The overhead arrows indicate that force and acceleration are vector quantities, which means they each have a magnitude and direction. Another example of a vector is velocity. The magnitude of a car’s velocity is its speed, but very often we also need to know the direction in which the car is travelling. In Newton’s first law, when we say that in the absence of forces a body moves at a constant velocity, this means that both the magnitude and direction of the velocity remain constant. When we want to refer only to the magnitude of a vector quantity, we drop the overhead arrow. For example, F is the magnitude of \vec{F} and $a = F/m$ means that the magnitude of the acceleration is given by the magnitude of the force divided by the mass.

We can arrange an experiment in which the same force is applied to two different masses. Equation (2.1) tells us that the acceleration of the larger mass will be less than the acceleration of the smaller mass. Thus mass is a measure of a body’s resistance to acceleration. More massive objects are harder to accelerate.

Force is measured in Newtons, which can be expressed in terms of other units as: $1 \text{ N} = 1 \text{ kg m/s}^2$. One Newton is the force required to accelerate a one kilogram (1 kg) mass at 1 m/s^2 . It is important to remember that physical quantities only have meaning when we specify units. For example, if someone asks you how old you are and you reply 240, they would think you’re crazy. However, if you said 240 months, they would probably convert that to 20 years, and think it just a little odd that you chose to measure your age in months instead of years. It is also essential to use consistent units throughout any calculation.

A common misconception is to think that the direction of an applied force is always the same as the direction of motion. We need to remember

that a net force acting on an object produces an *acceleration* in the same direction as the force, but the *velocity* of the object might be in a different direction. For example, suppose you are traveling in your car at a uniform speed, and then you apply the brakes. The force your brakes apply is in the opposite direction to motion, although your declining velocity is still in the original direction.

We have been discussing Newton's laws governing the motion of objects.³ Although we are all familiar with velocities and accelerations from our everyday experience, it is important to point out that when we say an object is moving, we need to specify what it is moving with respect to. This defines a "reference frame". For example, during dinner on an airplane, your food tray is motionless relative to your lap, although relative to the ground it is traveling as fast as the plane. We can call your lap a "frame of reference" (the one in which the tray is still) and the ground is another, different frame of reference (relative to this frame the tray is moving very fast). So, a reference frame is an object relative to which we measure the locations and motions of other objects.

An *inertial frame of reference* is a frame associated with an object that is not acted upon by any net force and is moving by inertia. Once we specify one inertial frame of reference, any other frame that is moving with a constant velocity relative to the chosen frame, is also an inertial frame of reference. For example, the room you are in now is an inertial frame of reference (approximately).⁴ Any train outside that is moving with a constant speed relative to the room is also an inertial reference frame. Newton's laws apply in all inertial frames of reference, thus any experiment you do in your room will yield the same results as the identical experiment performed by a friend on one of those trains.

2.2 Newtonian Gravity

Every day we experience the force of gravity. Gravity is an attractive force—it brings objects together. Every atom in our bodies is attracted to the Earth. Furthermore, every atom in the Earth is attracted to us. In fact any

³Newton also formulated a third law, which states that in every interaction between two bodies, the force the first body exerts on the second body is equal and opposite to the force the second body exerts on the first. If you push your friend facing you on an ice-rink, she will coast backwards, but so will you.

⁴The Earth is not exactly an inertial frame because of its rotation about its axis, which can be observed with a Foucault pendulum.

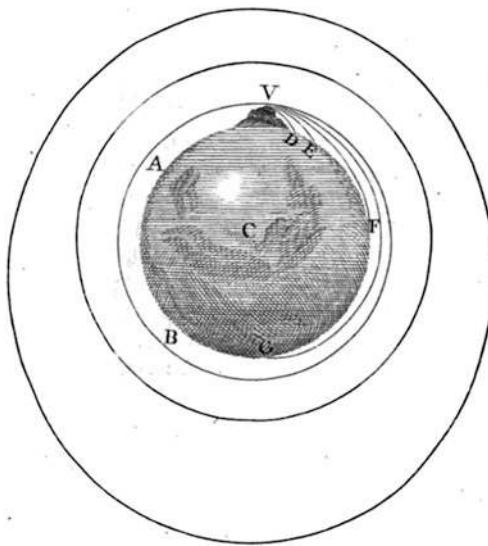


Fig. 2.2 Newton's thought experiment. Suppose a cannon is placed on top of a mountain and is fired with a moderate muzzle velocity. What will happen to the projectile? It will fall to Earth as shown at point *D*. If the muzzle velocity is increased it will fall a little farther away, as shown at points *E*, *F*, and *B*. Newton deduced that if the projectile is launched with progressively larger velocities, eventually, at just the right launch velocity, it will travel all the way around the Earth in a circular path, always falling in towards the Earth, but never reaching it, as indicated at *A*. Newton concluded that the Moon's orbit was of the same nature, with the Moon constantly falling toward the Earth. He also realized that if the launch velocity got higher, then elliptical orbits would be possible as shown. Credit Philosophiae Naturalis Principia Mathematica

two objects in the Universe exert a gravitational attraction on one another. Newton realized that the same kind of force responsible for an apple falling from a tree was also responsible for the revolution of the Moon around the Earth, and the Earth around the Sun (see Fig. 2.2). Thus his law of gravity is sometimes called *The Law of Universal Gravitation*, applying both to the Earthly and the heavenly realm.

Newton's law of gravity states that any two objects are attracted to one another with a force

$$F = \frac{GMm}{r^2} \quad (2.2)$$

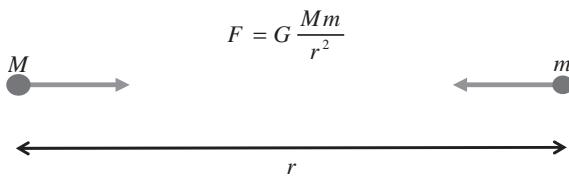


Fig. 2.3 Gravitational force of attraction between two point masses a distance r apart

where M and m are the masses of the two objects and r is the distance between them. The force acting on mass m is directed towards the mass M and vice versa (see Fig. 2.3). We have also introduced Newton's gravitational constant G , which has a measured value of $G = 6.67 \times 10^{-11} \text{ Nm}^2/\text{kg}^2$.

Newton's law of gravity is an "inverse square law", because in Eq. (2.2) the gravitational force is inversely proportional to the square of the distance between the two objects. For example, let M be the mass of the Earth and m the mass of the Moon. If the Moon were placed twice as far away from the Earth as its actual distance, then the Earth would exert a force of gravity on the Moon that is one quarter as strong as it currently is.

The masses in Eq. (2.2) are assumed to be "point masses"; that is, we assume that their sizes can be neglected, so we can imagine that each mass is located at a point. This is a good approximation for the Earth—Moon system: the sizes of the Earth and the Moon are much smaller than the distance between them, so they can be approximated as point masses located at their centers. Then, to calculate the gravitational force of attraction, we use the distance from the center of the Earth to the center of the Moon. The same logic applies to the Earth orbiting the Sun.

Furthermore, Newton proved the "shell theorem", which states two important facts: (1) A uniform spherical shell of matter attracts an outside object as if all of the shell's mass were concentrated at its center. This applies to any uniform spherically symmetric object, like a solid sphere, since the object can be thought of as consisting of shells. (2) The gravitational force exerted on an object that is *inside* a uniform spherical shell of matter is zero. This result is surprising. The object doesn't even have to be at the center of the spherical shell—it can be anywhere inside the shell, and it will still feel no force.⁵

⁵To prove the shell theorem, Newton represented the shell as consisting of a large number of point masses and added together the forces produced by all of these masses. He had to invent calculus to perform this calculation!

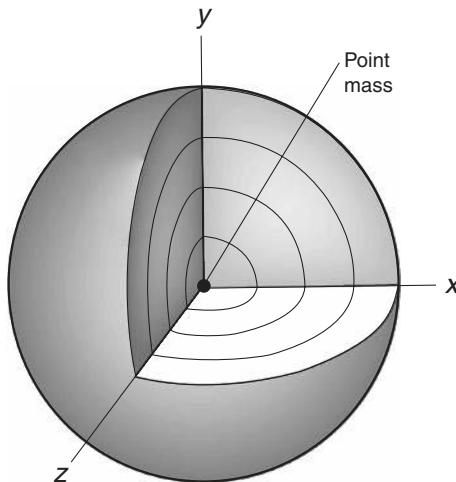


Fig. 2.4 Sphere divided into concentric shells. Each shell acts as if its mass is located at the center

To find the force of gravity acting on a small object near the surface of the Earth, we can imagine that the Earth (which is nearly spherical) is composed of a large number of thin concentric shells. Each shell will act as if all its mass is localized at the center, so the overall effect will be as if the entire mass of the Earth is localized at its center. Note that we do not have to assume that the mass density is uniform throughout the volume: each individual shell must have a uniform density, but the density can vary from one shell to the next. (In fact, the density of Earth is much greater near the center than near the surface.) (Fig. 2.4).

2.3 Acceleration of Free Fall

In everyday terms we often confuse weight and mass. When we get on a scale, we measure our weight—this is the force of gravity which pulls us towards the center of the Earth. For a small object of mass m near the surface of the Earth, the weight is given by

$$F = \frac{GM_E m}{r_E^2} \quad (2.3)$$

where $M_E = 6 \times 10^{24}$ kg is the Earth's mass, and $r_E = 6.4 \times 10^6$ m is its radius. On the Moon, we would weigh about 1/6 of our Earthly weight, even though our bodies would have the exact same amount of mass. The force of gravity on any object on the surface of the Moon is weaker than the gravitational force on that same object on the surface of the Earth. This is because the Earth has so much more mass than the Moon, that $(M/r^2)_{\text{Earth}} > (M/r^2)_{\text{Moon}}$, despite the fact that the Earth's radius is larger than the Moon's radius.

Now let's consider what happens if we have an object of mass m , close to the surface of the Earth, and we let it go. It will fall with acceleration $a = F/m$, which becomes (using Eq. (2.3) for F),

$$a = \frac{GM_E m}{r_E^2} \frac{1}{m} = \frac{GM_E}{r_E^2}. \quad (2.4)$$

This does not depend on the mass m , which means that all bodies close to the surface of the Earth fall with the same acceleration, independent of their mass (as long as we ignore air resistance). This remarkable fact was established by Galileo. The acceleration of free fall is denoted by the letter g ; its measured value is $g = 9.8 \text{ m/s}^2$. So, if we drop any object off a building, it will fall down with a velocity which increases by 9.8 m/s every second. Thus, after the first second the object will have a velocity of 9.8 m/s ; after the second, it will have a velocity of 19.6 m/s and so on (assuming the object is simply let go, with an initial velocity of zero). From Eq. (2.4) we see that

$$g = \frac{GM_E}{r_E^2} \quad (2.5)$$

Substituting the values of M_E , r_E and G (Newton's constant), you can verify that indeed $g = 9.8 \text{ m/s}^2$.

2.4 Circular Motion and Planetary Orbits

Velocity characterizes how fast the position of a body changes with time, and acceleration is the rate of change of velocity with time. When we travel at a constant speed and in a constant direction, our acceleration is zero. What happens if we travel around a large circular track at a constant speed? Do we accelerate? Yes, we do. Even though we maintain a strictly uniform speed, we constantly have to change the direction in which we are traveling. This

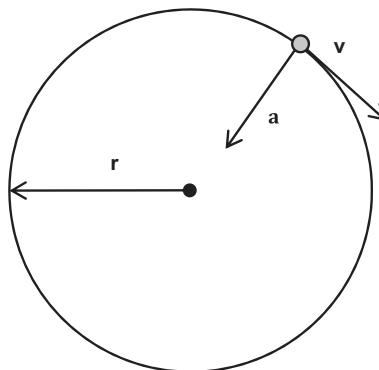


Fig. 2.5 Direction of acceleration for uniform circular motion

change in direction indicates that there is an acceleration. For uniform circular motion, the magnitude of the acceleration is

$$a = \frac{v^2}{r} \quad (2.6)$$

where v is the speed of the object undergoing the motion and r is the radius of the circle. The direction of the acceleration is radially inward, towards the center of the circle (see Fig. 2.5); it is called the centripetal acceleration.⁶ If you have ever swirled an object attached to a string above your head, you know that the tension in the string keeps the object from flying off at a tangent to its orbit. The string thus provides a force directed towards your hand that results in the object undergoing centripetal acceleration.

Newton showed in his Principia that the inverse square law implies that celestial bodies like planets and comets should move in elliptical orbits, in agreement with Kepler. While comets often move in highly eccentric orbits, for planets the two focal points of the ellipse almost coincide, so the orbit is approximately a circle. If the radius of the orbit is r , its circumference is $2\pi r$, and the velocity of the planet is

$$v = \frac{2\pi r}{T} \quad (2.7)$$

⁶The derivation of this formula relies on some simple geometry and can be found in any basic physics textbook.

where T is the time it takes to complete one revolution.

We can now apply what we have learned about Newton's laws to weigh the Sun. Let's do it!

We know that the force keeping the Earth in motion around the Sun is gravity; thus Eq. (2.2) holds, with m the Earth's mass, and M the Sun's mass. We also know that to a good approximation the Earth orbits the Sun with uniform velocity in a circle and thus undergoes centripetal acceleration (gravity is the force responsible for this centripetal acceleration). Then, substituting Eq. (2.6) into Eq. (2.1) and equating this with Eq. (2.2) we find $F = m\frac{v^2}{r} = GmM/r^2$.

Rearranging, the Sun's mass is given by

$$M = v^2 r / G \approx 2 \times 10^{30} \text{ kg}, \quad (2.8)$$

where the Earth's orbital velocity $v \approx 30 \text{ km/s}$ can be calculated from Eq. (2.7) using our knowledge that it takes the Earth one year ($T \approx 3 \times 10^7 \text{ s}$) to complete one orbit at a distance of $r \approx 1.5 \times 10^8 \text{ km}$. This method is often used in astronomy to measure the masses of stars, galaxies, and even clusters of galaxies.

2.5 Energy Conservation and Escape Velocity

Energy is nature's ultimate currency—it comes in several different forms, and can be converted from one form to another. For example, to launch a rocket into space, chemical energy must be converted into kinetic energy of motion. In general, the conservation of energy is one of the most fundamental laws of nature.

Here we will focus on mechanical energy. Mechanical energy can be divided into two types: kinetic energy and potential energy. Kinetic energy is the energy an object has by virtue of its motion. An object of mass m traveling with speed v has kinetic energy

$$K = \frac{1}{2}mv^2 \quad (2.9)$$

Potential energy is the energy a system has due to interactions between its parts. It can be thought of as stored energy that has the capacity to be unleashed and turned into kinetic energy. There is no universal formula for the potential energy; it depends on the kind of interaction. In the case of gravitational interaction between two spherical masses, it is given by

$$U = -\frac{GMm}{r} \quad (2.10)$$

where r is the distance between the centers of the spheres.⁷ If there are more than two masses, one simply has to add the potential energies for all pairs.

For a small object close to the Earth's surface, the potential energy (Eq. 2.10) can be approximated by the following useful formula:

$$U = mgh + \text{const} \quad (2.11)$$

Here, m is the object's mass, h is its height above the ground ($h = r - r_E$ where r is its distance from the center of the Earth, and r_E is the Earth's radius), and $g = 9.8 \text{ m/s}^2$ is the gravitational acceleration close to the Earth's surface.

The constant in Eq. (21.1) is $-GM_E m / r_E$, where M_E is the mass of the Earth. Such constant additions to the energy are unimportant for most purposes and are often omitted.

The total energy of the system is the sum of its kinetic and potential energies,

$$E = K + U \quad (2.12)$$

In an isolated system, to which no external forces are applied, the total energy is conserved—that is, it does not change with time. This is an immensely useful property, which makes the solution of many problems much easier than it would otherwise be.

A ball of mass m in a frictionless U-shaped track provides a classic example of the interplay between potential and kinetic energy (see Fig. 2.6). Let's place the ball on the left arm of the track, so that it is at a height h above the bottom of the track. We will let go of the ball, and it will start rolling. The ball's initial speed will be zero, but as it rolls down the track it picks up speed, attaining its maximum velocity at the bottom of the track (technically the ball has a rotational velocity in addition to its translational velocity, which we will ignore here for clarity. In other words, we will treat the ball as though it is "sliding" down the track). It will then rise up the right arm of

⁷This formula can also be used for a small object (like a human) interacting with a large spherical body (like the Earth). In this case, the small object does not have to be spherical, and the distance r is the distance from any point in the object to the Earth's center.

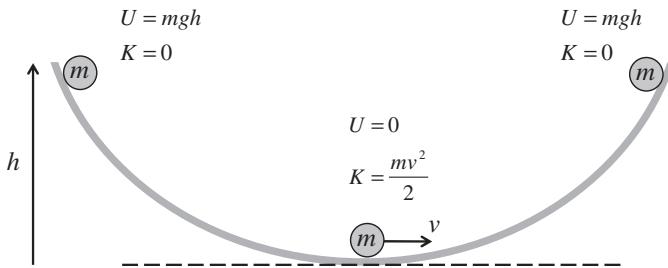


Fig. 2.6 At the *top* of the track, all the energy is in the form of potential energy. At the *bottom*, all potential energy has been converted to kinetic energy, and the ball has its maximum velocity. On its way up or down the track there is a mix of potential and kinetic energy, but the total mechanical energy is the same at every point along the ball's trajectory

the “U”, until it reaches a maximum height and momentarily comes to rest before rolling down the right arm.⁸ How high up the right arm does it get?

The answer is easy. By conservation of energy, it must reach the same height h as it started with. When the ball is at its starting point it has no kinetic energy (it is released from rest), but it has gravitational potential energy $U = mgh$. When it reaches the maximum position on the right arm, it also has no kinetic energy, since it is momentarily at rest. So it must have the same amount of potential energy, thus it must reach the same height h .

How fast will the ball be moving at the bottom of the “U”? At the bottom, the ball has no potential energy because it has zero height above the reference ground level. Thus all its initial potential energy is converted into kinetic energy and we have $\frac{1}{2}mv^2 = mgh$, which can be solved to yield the velocity at the bottom if we know h . In fact, we can find the velocity of the ball at any point along its motion if we know the height at that point.

Similar interchange between kinetic and potential energies occurs when planets move around the Sun. The expression for the potential energy U to use in this case is Eq. (2.10). This formula can be a little tricky to deal with, so it is useful to consider a plot of U versus r , as shown in Fig. 2.7.

We see that U approaches a maximum value of zero, when two objects are separated by larger and larger distances. But because the sign of the potential energy is negative, the gravitational potential energy decreases as objects are brought closer to each other (watch out for the minus sign!). Hence the

⁸You probably remember what happens when you are pulled up and released on a swing: you start from rest, reach a maximum speed at the bottom of your trajectory and then slow down as you swing up, momentarily coming to rest before going backwards, and so on.

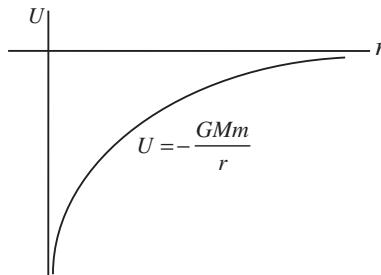


Fig. 2.7 Gravitational potential energy as a function of distance

kinetic energy should grow and the objects should move faster as they get closer. A planet moving along its elliptic orbit speeds up as it gets closer to the Sun and slows down as it gets further away.

The gravitational potential energy between two orbiting bodies can be thought of as a binding energy. The closer the two bodies are, the more negative is the potential energy, and thus we would have to work harder, or put in more energy, to separate them.

Since mechanical energy is the sum of kinetic and potential energy, it is possible for a pair of orbiting objects to have negative, zero or positive total mechanical energy. When the total energy is negative, it simply means that the system has less kinetic energy than the magnitude of the gravitational potential energy. This is in fact the case for all bound orbits, like the Earth-Moon system, the comets that orbit the Sun, or even the man-made satellites that orbit the Earth. Orbits with zero or positive total mechanical energy are said to be unbound. For example there are currently five spacecraft, Voyager 1 and 2, Pioneer 10 and 11, and the New Horizons Spacecraft, which are heading out of our Solar System on unbound orbits (or escape trajectories).

You might be wondering, how do we control whether a satellite we launch goes into orbit around the Earth, or goes off into the Solar System? The answer is very simple. There is a minimum initial speed, called the escape speed, with which the object must be launched in order for it to escape from the Earth. So, how do we calculate this escape speed? We use the principle of energy conservation and the fact that for the object to escape, its total mechanical energy must be greater than or equal to zero. Let's say we launch a spacecraft of mass m with speed v . Its total initial mechanical energy as it leaves the Earth is

$$E_i = \frac{1}{2}mv^2 - \frac{GM_E m}{r_E}, \quad (2.13)$$

which must equal its total final energy E_f when it has escaped.⁹ If we take the marginal case of zero total energy, the launch speed is by definition the escape velocity, and we have

$$\frac{1}{2}mv_{esc}^2 - \frac{GM_E m}{r_E} = 0. \quad (2.14)$$

This can be solved to find the escape speed

$$v_{esc} = \sqrt{\frac{2GM_E}{r_E}}. \quad (2.15)$$

Substituting in values for the Earth's mass and radius, we find $v_{esc} \approx 11.2 \text{ km/s}$. So if we launch a satellite with a speed slightly greater than 11.2 km/s, it will leave the Earth's gravitational clutches. If we launch it with less than this speed, it will fall back to Earth. And if we launch it at exactly the escape speed, it will barely escape, with its velocity getting smaller and smaller as it moves away and approaches zero in the limit.

Note that although we have derived the escape velocity for an object launched from the Earth, this formula holds in general, with the Earth's mass and radius replaced by whatever body you are considering. In later chapters we will apply similar considerations to the entire universe. Note also that the escape (or no escape) outcome depends only on the magnitude of the velocity, not on its direction.

2.6 Newtonian Cosmology

Newton's cosmological ideas developed during a correspondence with the Cambridge theologian Richard Bentley. Bentley was preparing to give public lectures titled "A confutation of atheism" and wrote to Newton asking him how his theory of gravity applied to the universe as a whole. During the winter of 1692–93, Newton sent a series of four letters to Bentley, in which he described a universe that is infinite and static: "The fixed stars, being equally spread out in all points of the heavens, cancel out their mutual pulls by opposite attractions."

⁹Note that the final energy is purely kinetic and must therefore be positive. This says that only objects with positive (or zero in the marginal case) total energy can escape.

However, as Newton was acutely aware, there is a problem with this line of reasoning. If a region of the universe has a slight excess of matter, then that region will begin to attract material from its surroundings. The region will become denser, and it will attract more and more matter. Thus a uniform distribution of stars is unstable due to gravity: it would be destroyed by an arbitrarily small perturbation. Newton's solution was to invoke a supernatural intervention, stating "...this frame of things could not always subsist without a divine power to conserve it."

2.7 Olbers' Paradox

What do you think the Sun would look like if it were at twice its present distance from the Earth? The total brightness of the Sun would be four times smaller because the brightness of an object decreases as the inverse square of the distance to the object.¹⁰ The area of the Sun's disc on the sky would also be four times smaller. This means that the brightness per unit area (called the surface brightness) remains the same. So what? Well, in an infinite universe that is uniformly sprinkled with stars, every line of sight should eventually hit a star, and each star should have roughly the same surface brightness as the Sun. This implies that the whole sky should be glowing with the same intensity as the Sun's surface. So why is the sky dark at night? This paradox is known as Olbers' paradox or the "dark night sky paradox". It indicates that Newton's picture of the universe cannot be right (Fig. 2.8).

An infinite static universe has other problems in addition to the gravitational instability and Olbers' paradox. We shall return to this issue in Chap. 5. For now, we just note that the problems of Newtonian cosmology give a foretaste of things to come: it is not so easy to come up with a cosmological model that makes any sense at all.

While Newton showed that his universal law of gravity could explain a vast scope of natural phenomena, he was at a loss to explain how it could be that the force of gravity acts instantaneously, between every pair of particles, across the vastness of space. This mysterious action-at-a-distance fueled Newton's critics. At the end of his Principia, Newton conceded: "*Thus far I have explained the phenomena of the heavens and our sea by the force of gravity, but I have not yet*

¹⁰This will be discussed in Chaps. 4 and 6.

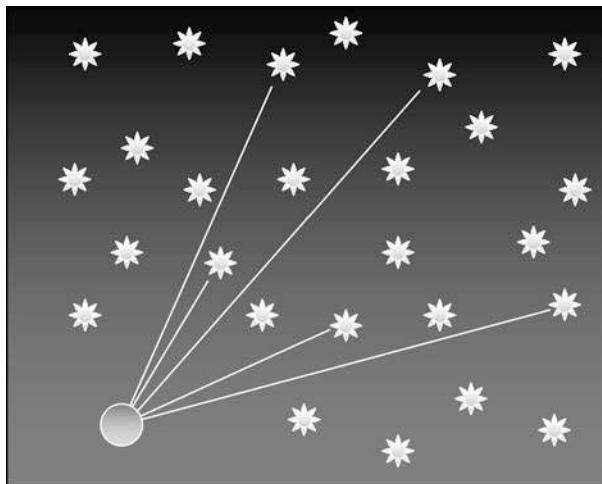


Fig. 2.8 Olbers' paradox. The entire sky should be as bright as the Sun. The astronomer Heinrich Wilhelm Olbers popularized the paradox, although he was not the first person to formulate it

assigned a cause to gravity.... I have not as yet been able to deduce from phenomena the reason for these properties of gravity and I do not feign hypotheses,” but he then went on to declare “And it is enough that gravity really exists and acts according to the laws that we have set forth and is sufficient to explain all the motions of the heavenly bodies and of our sea.” Despite these rumblings, Newton’s description of gravity held sway for two hundred years—until Einstein’s general theory of relativity revolutionized our understanding of gravity once again, as we shall later learn.

Summary

Newtonian mechanics forms the foundation for our understanding of the physical universe. We used Newton’s laws of motion and his universal law of gravity to explore planetary orbits, energy conservation and escape velocities. We then discussed Newton’s cosmological picture of an infinite, static universe uniformly filled with stars. Amongst other problems, this picture is incompatible with the observation of a dark night sky—this is known as Olbers’ paradox. The problems of Newton’s static universe give us an inkling of how difficult it is to develop a sensible cosmology.

Questions

1. Can you give an example, from everyday life (that is not mentioned in the text), where the force applied to an object is not in the direction of motion?
2. Is every frame of reference an inertial frame of reference?
3. In what ways are the laws of Nature, like Newton's laws, different from criminal laws?
4. If you apply the same force to two boxes, one of which is twice as heavy as the other, how will their accelerations compare? (Assume there is no friction)
5. What would happen to the force exerted by the Moon on the Earth if the Moon were placed twice as far away? And if it were brought into a third of its current distance? In which direction does the Earth pull on the Moon? In which direction does the Moon pull on the Earth?
6. Suppose you weigh 150 lbs. What would you weigh if the Earth was shrunk to half its current radius? (Assume that you and the Earth have the same mass before and after the contraction.)
7. Suppose we dig a tunnel radially through the Earth. If you weigh 150 lbs on the surface of the Earth, what would your weight be if you descend half way towards the Earth's center? (Assume that the Earth has uniform density throughout its volume. Also note that the volume of a sphere is given by $V = \frac{4}{3}\pi r^3$ where r is the radius, and that density is $\rho = M/V$ where M is mass, and V is volume.)
8. (a) Find the Earth-Moon distance given that it takes a radio signal 1.3 s to travel to the Moon. Note: radio signals travel at the speed of light which is denoted c and is approximately 3×10^8 m/s.
 (b) It takes roughly 27.3 days for the Moon to orbit the Earth (sidereal month). Calculate the velocity of the Moon around the Earth.
 (c) Use your results to help you calculate the mass of the Earth.
9. A space probe is launched from Earth with twice the escape velocity, $v = 2v_{esc}$. What will the velocity of the probe be when it gets far away from Earth? Express your answer in terms of v_{esc} .
10. A ball is released from rest at height h in a landscape shown in the figure. Assuming no friction, where will the ball reach its maximum speed? Indicate on the figure where it will come momentarily to rest (Fig. 2.9).
11. What is Olbers' paradox? Does it prove that the universe cannot be infinite? If not, what does it prove?

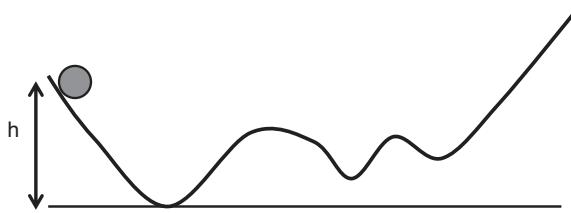


Fig. 2.9 Ball on a landscape

12. To explain Olbers' paradox, we argued that the brightness of a star and the area it subtends on the sky are both inversely proportional to the square of the distance to the star. Can you justify these statements?
13. Can you think of two ways to resolve Olbers' paradox?

3

Special Relativity

3.1 The Principle of Relativity

Although the word “relativity” is synonymous with Albert Einstein, Galileo Galilei was actually the first scientist to formulate what we now call “the principle of relativity”. In his famous *Dialogue* (1632), Galileo suggested to his readers that they should shut themselves up in the cabin of a ship, below the decks, so that they could not see what was going on outside. He argued that they would not be able to tell whether the ship was stationary or moving, as long as the motion was uniform (that is, at a constant speed) and in the same direction, without any turns. While the ship was the vehicle of choice in Galileo’s time, all of us can recall a similar experience in a train or an airplane. When the ride is smooth and you do not look outside, you cannot tell whether you are moving or not. Moreover, at this very moment you are moving with the Earth around the Sun at a speed of 30 km/s, and with the entire Solar System around the galactic center at about 230 km/s. These tremendous velocities have no effect whatsoever on anything we perceive here on Earth (Fig. 3.1).

The principle of relativity asserts that the laws of physics apply equally to all inertial observers. So if one inertial observer glides past another one, it is meaningless to ask which of the two observers is at rest and which is moving. There is no such thing as absolute rest or absolute motion: only relative motion has meaning (Fig. 3.2).

All of this was well understood and agreed upon since the time of Galileo and Newton, except for a rather mysterious problem that suddenly emerged

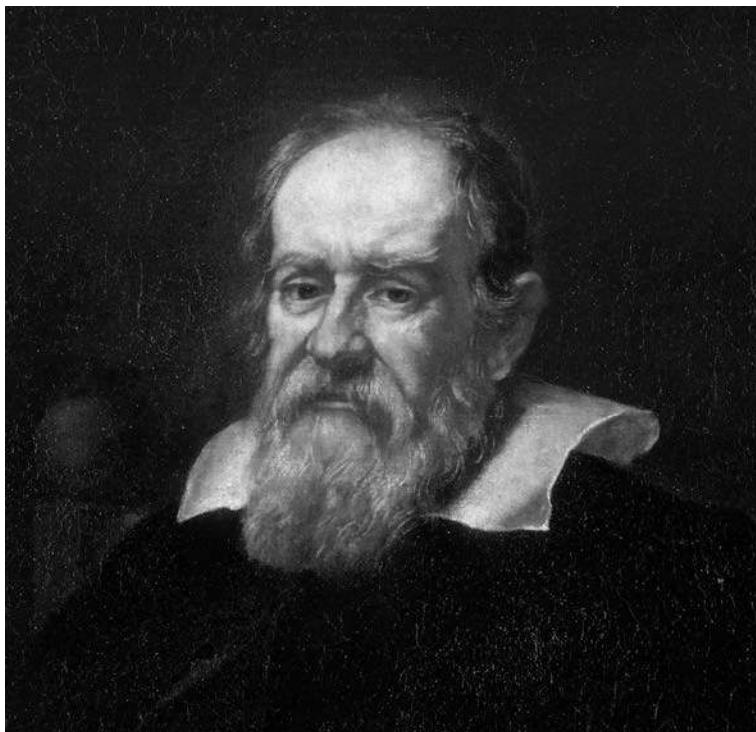


Fig. 3.1 Galileo (1564–1642) is often called the father of modern science. He experimented with balls rolling down inclined planes, which led him to deduce the law of inertia, and also the law of free fall, which states that all objects, regardless of their mass, fall to the earth with the same acceleration. Galileo was also the first person to point a telescope to the sky, revealing that there are thousands of stars invisible to the naked eye; that four moons orbit Jupiter; and that Venus goes through crescent and gibbous phases like the moon. These astronomical observations provided evidence for the Copernican heliocentric system, which he then endeavored to convince the world of in his book *Dialogue on the Two Chief World Systems, Copernican and Ptolemaic* (1632). The book ran contrary to the official dogma and was banned the following year by the church (The church allowed discussion of the Copernican system, as long as it was presented as a theory, not as fact. The true nature of the universe had to be deduced from Scriptures, not from astronomical observations). At age 70, Galileo was tried by the Inquisition, forced to recant, and then spent the rest of his life under house arrest

near the end of the 19th century. The problem involved the speed of light, which was measured to be about 300,000 km/s. The issue was, speed with respect to what? When we say that the speed of sound is 300 m/s, it is understood that sound propagates through the air to the observer from its source. Thus an observer, who is at rest with respect to the air, will measure this value for the speed of sound.

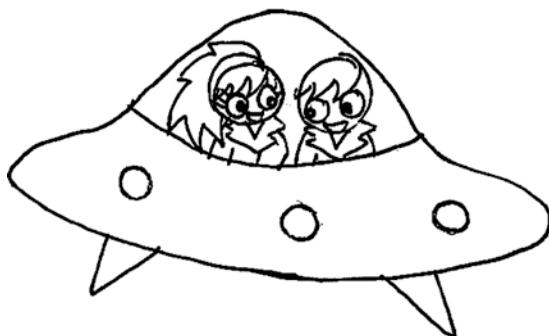


Fig. 3.2 Inertial observers move by inertia, unaffected by any forces. For example, an alien in a spaceship which has its engines turned off as it glides through interstellar space is an inertial observer. The spaceship is an inertial frame of reference. The passengers in Galileo's ship are also inertial observers, and the ship itself is an inertial frame of reference. Credit Natalie Perlov

The situation with light was expected to be very similar. Scientists were convinced there should be some substance through which light propagates; the so-called ether. Then, if we are at rest with respect to the ether, we should find that light propagates in all directions at the same speed of 300,000 km/s (see Fig. 3.3). But what if we are moving through the ether at a speed of, say, 100,000 km/s? Then a light pulse propagating in the same direction will only move at 200,000 km/s relative to us, while a pulse propagating in the opposite direction will speed away from us at 400,000 km/s. The speed of light in the directions orthogonal to our motion would remain unchanged at 300,000 km/s. Thus, in a space filled with ether, inertial observers who move at different speeds will no longer be equivalent. They will observe that the speed of light is direction-dependent, and they can use this to deduce their own speed with respect to the ether.

An ingenious apparatus for measuring the difference between the speeds of light in two orthogonal directions was designed by Albert Michelson of the Case School in Cleveland and Edward Morley of the neighboring Western Reserve College. They knew that if our speed through the ether was as high as 100,000 km/s, then the direction-dependence of the speed of light would have been noticed in earlier experiments. Thus, Michelson and Morley expected to measure a much smaller number—which required a much higher accuracy. They figured that, if nothing else, the Earth's speed through the ether should not be much smaller than the speed of its

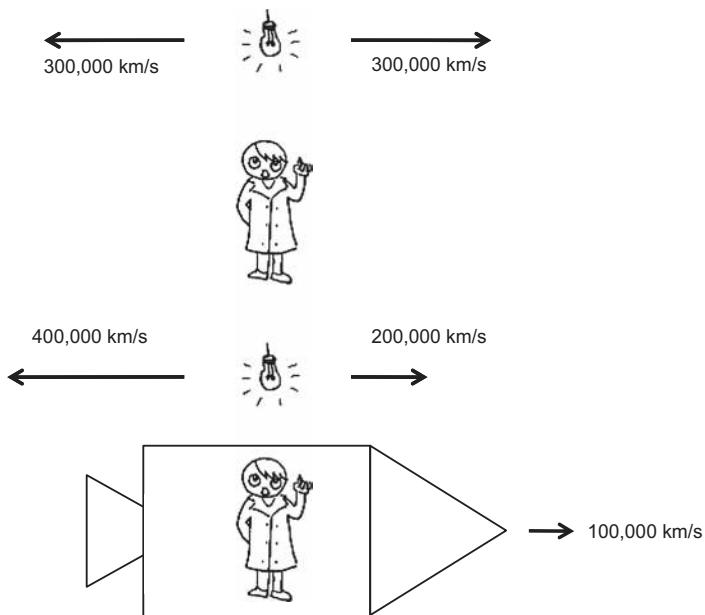


Fig. 3.3 An observer would measure a different value for the speed of light, depending on whether he is stationary or moving relative to the ether. Credit Natalie Perlov

revolution around the Sun, which is 30 km/s. Fortunately this was well within the measurement capabilities of their instrument.

Michelson and Morley performed their experiment in July of 1887. The result was startling: they detected no variation in the speed of light in different directions. None whatsoever. To ensure that the Earth's motion around the Sun at the time of the experiment was not accidentally compensated by the motion of the whole Solar System through the ether in the opposite direction, Michelson and Morley repeated the experiment six months later. By that time, the Earth completed half a revolution and was moving in the opposite direction, so if the two velocities initially cancelled, they must now add up, and the Earth should be moving through the ether at the speed of 60 km/s. But again, no effect was found.

A straightforward reading of their results suggested that the speed of light does not depend on the speed of the observer who measures it. This looked completely absurd and in conflict not only with Newtonian physics, but also

with common sense. So most physicists simply chose to ignore the results of the experiment.

The resolution of the paradox came in 1905 from a twenty six year old clerk working in the patent office in Bern, Switzerland. His name was Albert Einstein. He accepted the constancy of the speed of light as a fact and used it as a basis for a theory of stunning beauty. As for Newtonian physics and common sense, they had to go. So did the ether.

3.2 The Speed of Light and Electromagnetism

The truth is that the Michelson-Morley experiment played little role in convincing Einstein that the speed of light is the same for all observers. He had other reasons for postulating the constancy of the speed of light, which had to do with the theory of electromagnetism developed by James Clerk Maxwell in the mid-19th century (Fig. 3.4).

As a child, you probably enjoyed playing with magnets and making your hair “stand up” by bringing your freshly rubbed plastic ruler close by. At the time, you probably thought of magnetic and electric forces as two completely separate phenomena. But today, you may know something that even Kepler, Galileo and Newton did not: electricity and magnetism are two sides of the same “electromagnetic coin”.

The theory of electromagnetism describes the behavior of electric and magnetic fields, which are produced by static electric charges and by flowing electric charges, called currents (see Figs. 3.5 and 3.6). For example, the Earth’s magnetic field is produced by the electric currents flowing in its core. The field is present everywhere in space and becomes manifest when you hold out a compass. The compass needle will point in the direction of the field, and the field strength can be judged by how forcefully the needle swings in that direction. Similarly, an electric field permeates the space around bodies carrying positive or negative electric charges. The electric field causes attraction between opposite charges and repulsion between same-sign charges.

Maxwell expressed his theory in the form of eight equations that describe all electric and magnetic phenomena. The theory also made two very important predictions:

1. Oscillating electric and magnetic fields propagate through space as electromagnetic waves, and
2. The speed of an electromagnetic wave is 300,000 km/s.

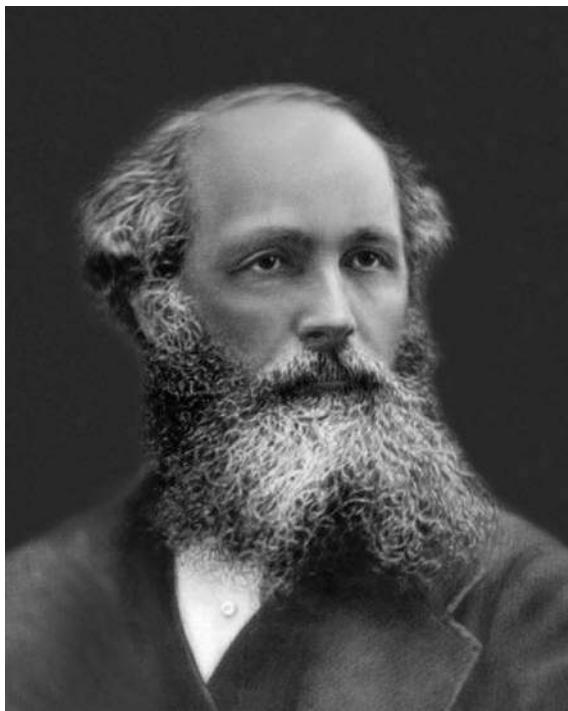


Fig. 3.4 James Clerk Maxwell (1831–1879) developed a unified description of electric and magnetic phenomena. He introduced the fundamental concept of a field, which now plays a central role in physics. Electric and magnetic fields propagate through space as electromagnetic waves, and Maxwell suggested that light must be a form of electromagnetic radiation. In the words of Nobel laureate Max Planck, he “achieved greatness unequalled.” And yet, Maxwell’s theory was largely ignored for more than 20 years after its publication. Maxwell was a shy and gentle person. He wrote poetry and felt a great affinity to animals. His writing was a model of clarity, but his conversation and lectures were rather confusing: he jumped from one subject to another, as his speech could not keep up with the pace of his thoughts. Maxwell died of cancer at the age of 48, well before his theory was widely accepted

Maxwell, of course, recognized this speed as the speed of light.¹ He thus came to a remarkable realization that light must be electromagnetic waves.

An electromagnetic wave is characterized by its *wavelength*, which can be defined as the distance from one “crest” of the wave to the next (see Fig. 3.7). Another useful characteristic is the *frequency*, defined as the number of crests

¹According to current measurements, the speed of light is 299792.458 km/s and it is convention to denote it by the letter c .

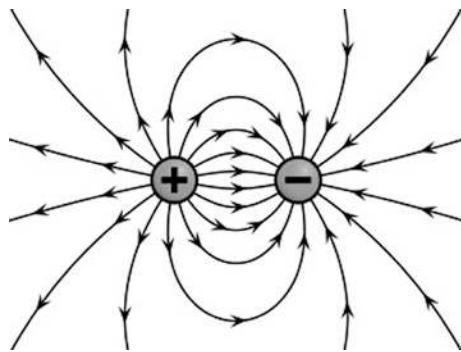


Fig. 3.5 Electric field produced by a positive and negative charge

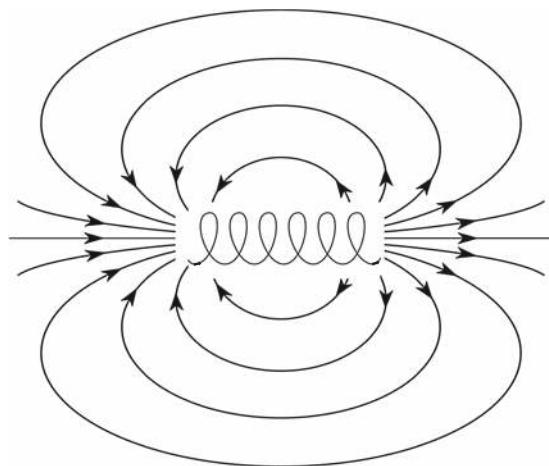


Fig. 3.6 Magnetic field produced by a current flowing in a coiled wire

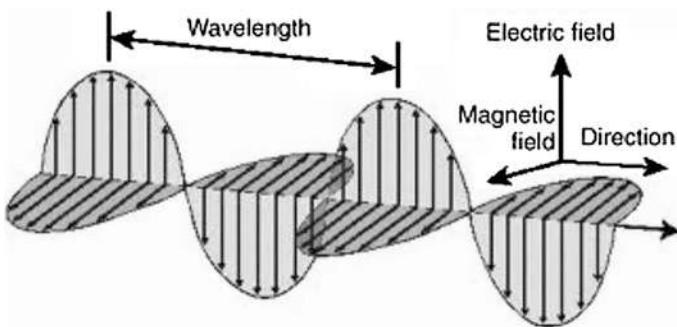


Fig. 3.7 An electromagnetic wave consists of electric and magnetic fields propagating through space at right angles to one another. Credit NASA

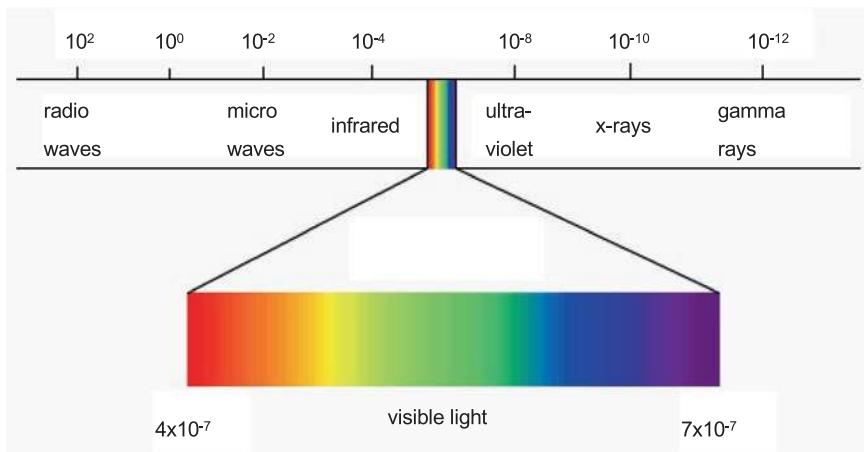


Fig. 3.8 The electromagnetic spectrum. All wavelengths are measured in meters here. For comparison, the width of a human hair is 8×10^{-5} m

passing through any given point per unit time (e.g. per second). The wavelength λ and the frequency f are related by a simple formula

$$f = \frac{c}{\lambda} \quad (3.1)$$

where c is the speed of light. Hence, the shorter the wavelength, the higher is the frequency.

It turns out that electromagnetic waves can propagate with many different wavelengths. Visible light corresponds to a narrow range of wavelengths in the full electromagnetic spectrum (see Fig. 3.8). The microwaves in your kitchen have longer wavelengths than visible light, while the X-rays at your

dentist's office have shorter wavelengths. Each type of electromagnetic radiation opens a unique window through which we can observe our radiant universe.

The theory of light as an electromagnetic wave is a pillar of classical physics. However, in some circumstances, such as when light interacts with atoms, it is necessary to adopt the viewpoint of quantum physics, which describes light as consisting of particles, called photons. In this modern picture, an ordinary light wave is thought to be composed of a huge number of photons that travel together. Each photon is electrically neutral and has a measurable energy and momentum. The energy of a photon is related to its wavelength. Blue light photons are more energetic than photons of the longer-wavelength red light. And gamma ray photons are much more energetic than the photons of visible light. For the rest of this chapter we can think of light as a classical electromagnetic wave.

3.3 Einstein's Postulates

In the early 1900s Einstein started to wonder what would happen if he followed Galileo's advice and locked himself up in the cabin of a ship (Fig. 3.9). Would he be able to detect the uniform motion of the ship by performing experiments with electric charges and currents? His intuition said "no". In other words, he believed that Maxwell's laws of electromagnetism are equally valid in both uniformly moving and stationary cabins. But then the speed of light should also be the same and it should equal 300,000 km/s, as required by Maxwell's theory.

In his celebrated 1905 paper *On the Electrodynamics of Moving Bodies*, where he first presented his special theory of relativity, Einstein postulated two fundamental assumptions:

- (1) The laws of physics are the same for all inertial observers. This is Galileo's *principle of relativity*.
- (2) The speed of light measured by all inertial observers is the same.

It follows from the second postulate that there is no ether²: electromagnetic waves propagate through empty space. Note also that the second postulate follows from the first if Maxwell's theory is included among the laws of physics.

²Otherwise the speed of light would vary depending on the observer's speed relative to the ether.

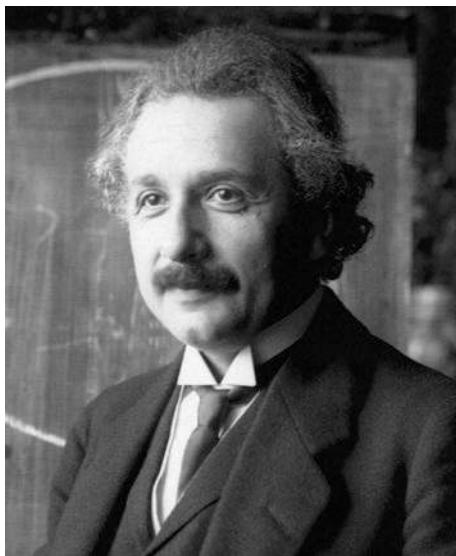


Fig. 3.9 Albert Einstein's career in physics had a bumpy start. After graduating from Zurich Polytechnic, he got no job offers in academia and thought himself lucky to find a job as a clerk at the patent office in Berne. On the positive side, the job was not very demanding and left plenty of time for Einstein's research and other intellectual pursuits. He spent evenings with friends, playing violin, reading philosophy and discussing his physics ideas. It was during this period that Einstein had his annus mirabilis (1905), publishing in the same year his special theory of relativity and his groundbreaking work on quantum mechanics. From this point on, Einstein's career took off, and within a few years he had job offers from several major universities in Europe. In 1919, when the light deflection predicted by general relativity was confirmed by observation, his fame reached the level of a pop star, and Einstein became a household name. He quipped that "To punish me for my contempt for authority, fate made me an authority myself." In his later years, Einstein often invoked God to express his philosophical views. He said, for example, "God does not play dice with the universe", expressing his doubts about the probabilistic interpretation of quantum mechanics, and "What really interests me is whether God had any choice in the creation of the world." He clarified, however, that "I believe in Spinoza's God who reveals himself in the harmony of what exists, not in a God who concerns himself with the fates and actions of human beings." Credit Albert Einstein lecturing in Vienna, 1921. Photo by Ferdinand Schmutzler

From these two simple assumptions, Einstein led us directly to the relativity of simultaneity, time dilation, length contraction and the equivalence between mass and energy. We will discuss each one of these in turn.

3.4 Simultaneity

Einstein liked to perform “thought experiments”—he would think up a hypothetical experiment, and then deduce the results through logical reasoning instead of making measurements. One of his famous thought experiments involves two observers, one in a moving train and the other stationary on the platform.

Suppose Jane is standing in the middle of a moving train car, holding the switch of an overhead light bulb, which is initially turned off. Now she turns it on, and light propagates towards the front and back walls of the car. Since she is in the middle, light reaches the two walls at the same time. At least from her point of view (Fig. 3.10).

Now suppose Ben is observing this experiment standing on the platform. He sees Jane turn on the bulb, and light propagates in both directions at the same speed, which is of course the speed of light. But the back wall of the car is moving toward the approaching beam of light, while the front wall is moving away from the beam, so Ben will see light reach the back wall before it reaches the front wall. The conclusion is that *the simultaneity of events has no absolute meaning; it depends on the state of motion of the observer*.

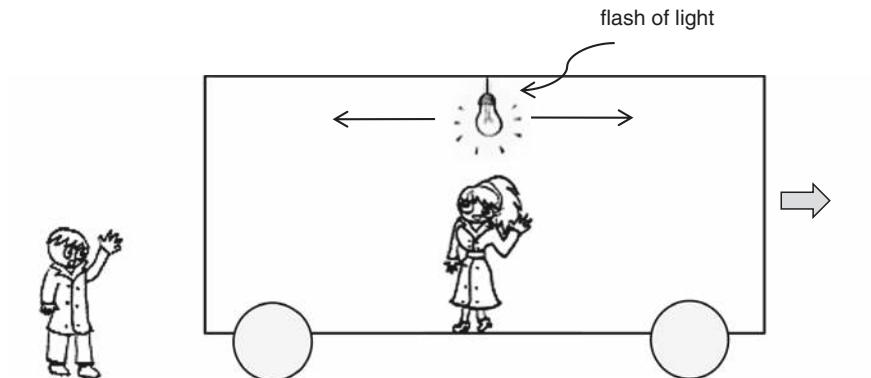


Fig. 3.10 The relativity of simultaneity. Two events that occur at the same time from the point of view of the observer in a moving train, occur at different times from the point of view of the observer on the platform. The arrival of light at the front, or the back wall of the car is an “event”. Credit Natalie Perlov

3.5 Time Dilation

In our everyday experience, a second is a second, regardless of whether we are waiting in line to board a plane, or if we are already in the air cruising at hundreds of kilometers per hour. However, according to Einstein, this obvious “fact” is not true. More precisely, the principle of relativity, and the constancy of the speed of light, lead inexorably to the following conclusion: *Moving clocks run slower as seen by an observer at rest.* Once again, this can be illustrated with the aid of our friends Jane and Ben.

Suppose now that Jane, who is in a train moving with speed v , is equipped with a light source, located at the floor of her train car, and a mirror at a distance L directly above the source (see Fig. 3.11a). A push of a button sends a light pulse from the source to the mirror and back again. This round trip takes time $t_0 = 2L/c$ (since light travels at a constant speed c), as measured by Jane.

Now, Ben, who is standing on the platform, sees the light pulse travel along a diagonal path (see Fig. 3.11b). Since the diagonal path length D is

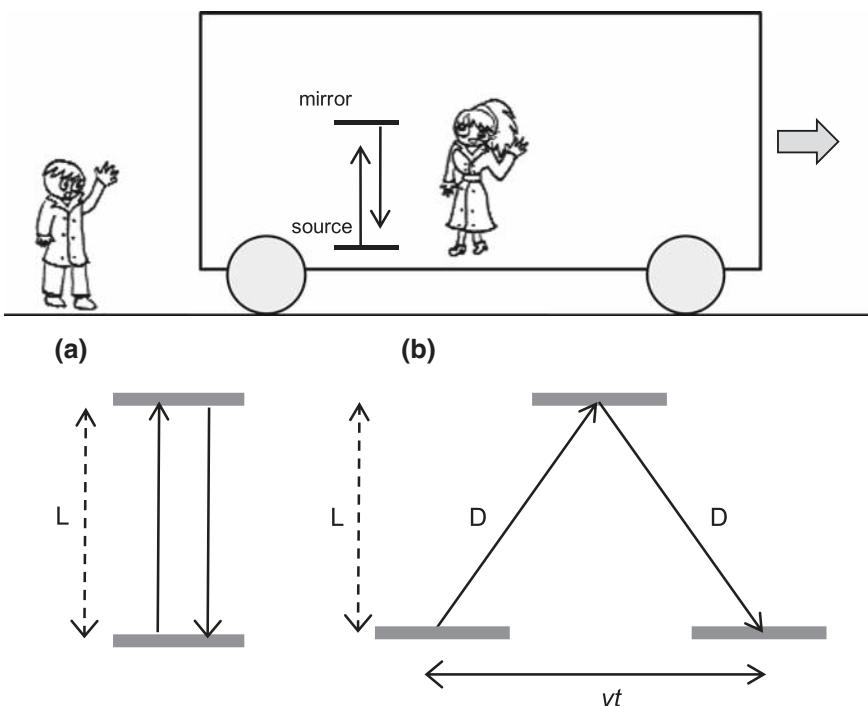


Fig. 3.11 Light pulse round trip: **a** Jane’s view. **b** Ben’s view. The light pulse travels a distance of $2D$, and the train travels a distance of vt . Credit Natalie Perlov

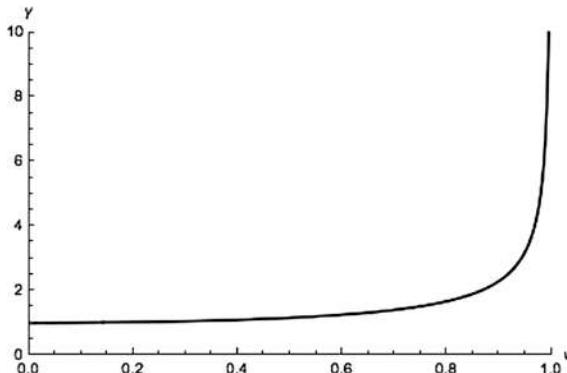


Fig. 3.12 The Lorentz factor γ as a function of velocity v (measured in units where $c = 1$)

longer than the distance L , he will find that it takes a longer time $t = 2D/c$ for the pulse to complete the round trip. Thus the time interval between the same two events (emission and return of the light pulse) is different when measured by different observers.

In particular, the time intervals that Ben and Jane will measure are related by the time dilation formula³

$$t = \gamma t_0 \quad (3.2)$$

The relativistic factor γ (also called the Lorentz factor, after Hendrik Lorentz who first introduced it) is defined as

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}} \quad (3.3)$$

The factor γ equals 1 when the relative velocity between two frames is zero, $v = 0$. Otherwise, it is greater than 1 and becomes arbitrarily large as the speed v approaches the speed of light.

Consider the plot of γ as a function of increasing velocity v (see Fig. 3.12). At first, γ grows very slowly. For example, if Jane's train moves at 300 km/h, then $\gamma \approx 1 + 10^{-13}$. This means that the time dilation Ben will measure amounts to about 1 s in a million years. This is tiny. You need to go at a sizeable fraction of the speed of light before γ differs appreciably from 1. For example, at 99.5% of the speed of light, time in the moving frame would go

³The derivation of this equation is presented at the end of the chapter.

slower by a factor of 10 compared to the time on the platform ($\gamma = 10$). If Jane flew away in a spaceship at that speed, then from the point of view of us here on Earth her life expectancy would be about 800 years. But since all processes in her body and brain would be slowed down by a factor of 10, she would not necessarily accomplish much more in her lifetime than a typical earthling. In Jane's reference frame, time would flow normally in the spaceship, but would slow down on Earth by a factor of 10.

You might think it should be possible to tell who has aged less by simply comparing Jane's age with that of her friends on Earth, at some moment of time. However, simultaneity is relative, and different observers will have different ideas about "the same moment of time". Jane might decide to resolve this issue once and for all, by turning her spaceship around and heading back to Earth. Alas, she will find that the earthlings were right. If she went away for 10 years, 100 years will have passed on Earth. But why? If motion is truly relative, why can't we think of Jane as being stationary with the Earth receding? As Jane turns on the engines and reverses the velocity of her ship, she is no longer an inertial observer, and thus the laws of special relativity do not apply in her reference frame.

3.6 Length Contraction

Once you know that time and simultaneity are relative, you will probably not be shocked to learn that space is relative as well. And indeed, Einstein's thought experiments demonstrated that moving objects contract in the direction of motion.

If L_0 is the length of an object at rest, and if the object then moves past an observer at speed v , the observer will measure the length of the object to be⁴

$$L = L_0/\gamma \quad (3.4)$$

where γ is the Lorentz factor defined in Eq. (3.3).

If Jane measures the length of her spaceship to be 50 m and she flies by the Earth at 99.5% of the speed of light, then we would find that her ship is only 5 m long. On the other hand, she will find the Earth and all its inhabitants to be squeezed tenfold in the direction of her motion (Fig. 3.13).

⁴The derivation of length contraction can be found in most basic undergraduate physics textbooks.

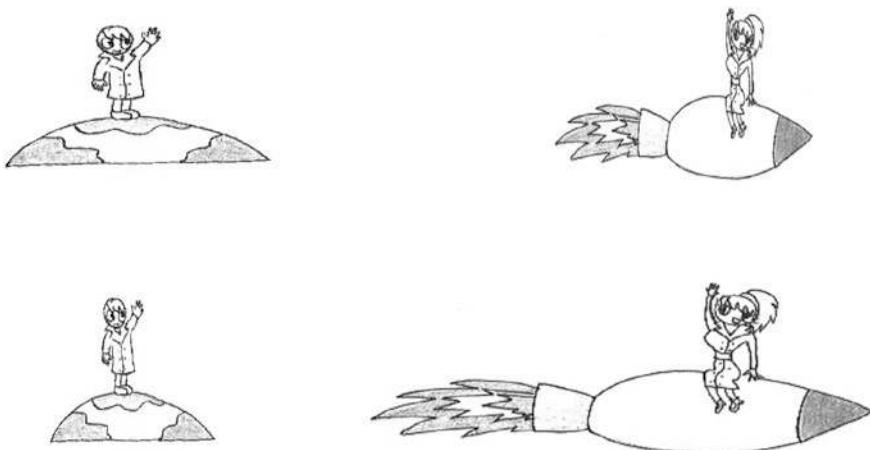


Fig. 3.13 Length contraction. To an observer on Earth, Jane and her spaceship are contracted in the direction of motion. From Jane's point of view, the Earth and its inhabitants are contracted. Credit Natalie Perlov

3.6.1 Speeding Muons

As intellectually chic as it may sound to talk of time dilation, length contraction, and so on, one can't help but wonder if these phenomena have any measurable effects here on Earth. Well, they do!

We are all familiar with the electron, but may be somewhat less familiar with its big cousin, the muon, which is about 209 times more massive. The muon is also negatively charged, but unlike the electron, it is unstable and decays into other kinds of particles in a mere 2.2×10^{-6} s (or $2.2 \mu\text{s}$), when at rest. Let's now imagine we have a laboratory that is 4 km long, with a muon emitter on one end and a muon detector on the other. We will also assume our muons are emitted with a speed of 99.5% that of light. *How far does a muon travel before decaying into other particles?*

"Non-relativistic" answer: about 0.66 km. So would you expect the detector at the end of your lab to detect the muon? No you wouldn't—it should decay before it has a chance to traverse the length of the lab. However, to the surprise of "non-relativistic experimenters", such muons are in fact detected. The mystery is resolved when we include the effects of time dilation.

According to you, experimenting in the lab, the muon is moving and thus its clock slows down by a factor of $\gamma = 10$. So while the muon "thinks" it has lived $2.2 \mu\text{s}$, you see it living for $10 \times 2.2 = 22 \mu\text{s}$. During this time interval, the muon can actually traverse a distance of 6.6 km—comfortably

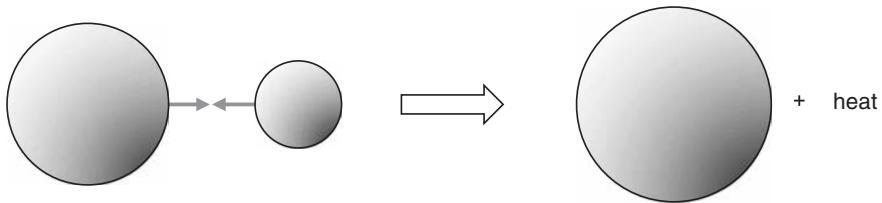


Fig. 3.14 Schematic illustration of nuclear fusion. Two nuclei combine to produce a new, larger nucleus, whose mass is smaller than the sum of the masses of its constituents. The difference in mass is given off in the form of heat (and light) energy

reaching the end of the lab and the detector! Such an extended lifetime of rapidly moving particles is now routinely observed in particle physics experiments.

But what would the muon conclude if it could think? According to the muon, it lives for $2.2 \mu\text{s}$, and thus expects to travel 0.66 km before decaying. However, it sees a lab rushing by at 99.5% the speed of light. So instead of the lab being 4 km long, the muon thinks the lab is only $4/10 = 0.4$ km! So to the muon, there is no doubt that it will make it to the end of the lab before decaying.

So, both the muon and the experimenter in the lab will agree that the muon should in fact be detected, although to the observer this is due to time dilation of the moving muon, and from the muon's frame of reference, it is due to length contraction of the moving lab.

There are many other practical applications of special relativity, which you could google if you feel incurably curious. Don't forget to find out how time dilation affects GPS satellites...

3.7 $E = mc^2$

A few months after Einstein published his first paper on relativity, he mined yet another gem from the two postulates: energy and mass are related in a fundamental way. It took another two years before he arrived at his famous equation $E = mc^2$, which he boldly and correctly interpreted to describe a complete equivalence between mass and energy. Loosely speaking, *mass can be converted into energy, and energy can be converted into mass*. An important example of mass being converted into energy occurs during nuclear reactions in the cores of stars (see Fig. 3.14).

For a physical object moving as a whole with velocity v , Einstein's relativistic energy relation can be expressed as

$$E = mc^2 = \gamma m_0 c^2 \quad (3.5)$$

where γ is the Lorentz factor, and m_0 is the mass of the object at rest, called its *rest mass*. For velocities much smaller than the speed of light, this formula is well approximated by $E = m_0 c^2 + 1/2 m_0 v^2$. Here, the first term is called the rest energy of the object and the second is its Newtonian kinetic energy.

Notice that the energy for an object at rest does not vanish—it is the rest mass m_0 (times c^2). The idea that an object at rest has energy by virtue of the fact that it has mass, was a totally new concept when Einstein introduced it. Another important feature of the relativistic energy is that it tends to infinity when the velocity of an object approaches the speed of light. It takes increasingly large amounts of energy to bring the speed of the object closer and closer to the speed of light, but the limit $v = c$ can never be reached. Thus, the speed of light is the absolute speed limit in the Universe.⁵

3.8 From Space and Time to Spacetime

We say that space is 3-dimensional because it takes three numbers to indicate a location in space. For example, you could arrange to meet someone in a restaurant on the 7th floor in the building at the corner of 16th Street and 5th Avenue. But what if you forget to specify *when* you want to meet? Clearly including the time of your proposed meeting is just as important as the location. Thus we can think of time as a 4th dimension.

When we combine our notions of space and time, we speak of *spacetime*. Since there are three spatial dimensions and one time dimension in our universe, we say spacetime is 4-dimensional. In Newtonian physics such a combination is artificial because space and time are completely independent. But in Einstein's theory it is very natural, so space and time are routinely depicted together in a spacetime diagram, as shown in Fig. 3.15. An *event* is a point in spacetime. It can be specified by four pieces of information, or four coordinates (t, x, y, z) , where the t stands for time and the other coordinates represent the location in space. The history of a point-like object is represented by a line, known as its *worldline*. The worldline tells us where in space the object is, for each moment in time. What do you think the worldline of an object at rest looks like?

⁵This limit is attained only by particles with a zero rest mass, like the photon.

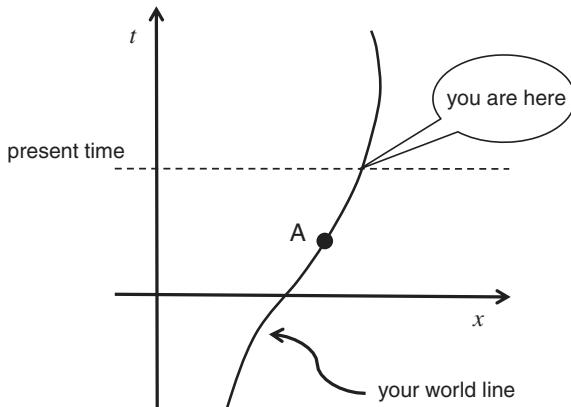


Fig. 3.15 Spacetime diagram showing time on the *vertical axis* and one spatial dimension on the *horizontal axis*. It is conventional to draw only one spatial dimension in the spacetime diagram because we can't draw all three spatial dimensions plus time. The present time is indicated by a constant-time “slice” on this diagram. Also shown is the history of a point-like version of yourself—your worldline—and some event A which took place in your past

If all of spacetime was laid out in front of you, then you would know everything about the past, present and future of the Universe. You would be able to follow the worldline of each particle and determine its location at any moment of time. A “moment of time” is a 3-dimensional slice through the 4-dimensional spacetime. All events on this slice are simultaneous from the point of view of a certain observer. Another observer, with a different notion of simultaneity, will draw a different slice. Although these 3-dimensional “snapshots” of the Universe may look different from one another, the underlying 4-dimensional spacetime is the same.

The notion of 4-dimensional spacetime was championed by Einstein’s former mathematics professor, Hermann Minkowski, who also uncovered its deep geometric structure. In a lecture at the University of Gottingen, Minkowski proclaimed that “Henceforth space by itself and time by itself are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality” (Fig. 3.16).

Minkowski’s insight can be explained by analogy with Euclidean geometry on a plane. Suppose we have two points on a plane that are connected by a straight line segment. We can characterize this segment by its projections on two orthogonal axes, x and y . The squared length of the seg-



Fig. 3.16 Hermann Minkowski was the son of Lithuanian Jews; he converted to Christianity to improve his job prospects. He taught Einstein mathematics at Zurich Polytechnic Institute. As a student, Einstein did not think much of Minkowski's lectures, while Minkowski remembered Einstein as a "lazy dog" and did not expect him to produce anything worthwhile. To Minkowski's credit, he changed his mind quickly after reading Einstein's 1905 paper. He pioneered the concept of a four-dimensional spacetime and used it to develop a geometric formulation of special relativity. Einstein was still unimpressed with his former professor and thought that Minkowski's mathematical luster only obscured the physical meaning of his theory. But soon it was Einstein's turn to change his mind. Minkowski's four-dimensional spacetime was indispensable for the construction of Einstein's theory of gravitation

ment equals $\Delta x^2 + \Delta y^2$, by the Pythagorean Theorem.⁶ We can now choose a different set of axes, rotated with respect to the first. The projections of our segment on the new axes will then be different, but the sum of their squares will be the same, and will remain equal to the length squared of the segment.

Minkowski realized that the situation in Einstein's theory is similar. Given two events, their separations in space and time will differ for different observers. But this is only because the events are being projected on different space

⁶Physicists often use the Greek letter Δ to denote the change in a given quantity. For example, in the text above, Δx means the change in the value of x from one end of the line segment to the other end.

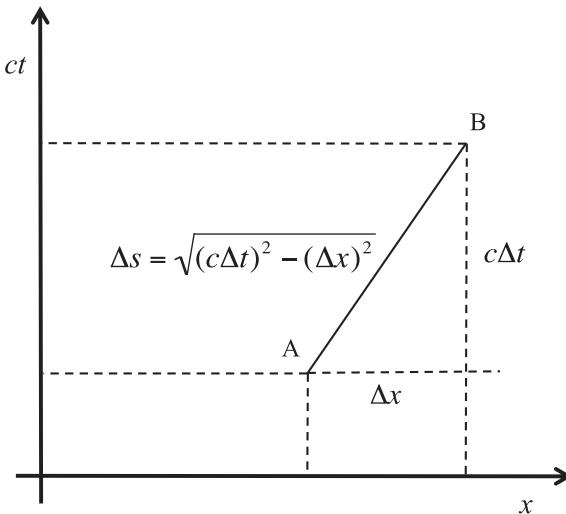


Fig. 3.17 Observers in different inertial frames of reference would measure the same spacetime interval between A and B , but they would get different values for the time elapsed and the spatial distance between the events

and time axes. Minkowski found that all observers will agree on a specific quantity called the *spacetime interval*. Consider events A and B marked in Fig. 3.17. Their space and time separations are denoted by Δx and Δt , respectively. The square of the spacetime interval can then be expressed as

$$(\Delta s)^2 = (c\Delta t)^2 - (\Delta x)^2 \quad (3.6)$$

Note that, except for the minus sign, this is very similar to the Pythagorean theorem, especially if we use units with $c = 1$ (see paragraph below). Since all inertial observers will agree on the value of the spacetime interval, we say that the spacetime interval is *invariant*. Note also that for a particle traveling at the speed of light (like a photon) the spacetime interval between any two events on its worldline is zero (can you show this?).

Whenever you graph physical quantities, it is crucial to know what units you are using. It turns out that for spacetime diagrams it is illuminating to use units in which the speed of light is $c = 1$. For example, time can be measured in years and spatial distances in *light years*. A light year is the distance light travels in one year, which is roughly 10^{13} km (can you show how we calculate this?). Then $c = 1$, since light travels a distance of one light year per year. When we plot the worldline of a light beam on a spacetime diagram using years and light years on the vertical and horizontal axes, respectively, we get a straight line at an angle of 45° to the axes, as shown in Fig. 3.18.

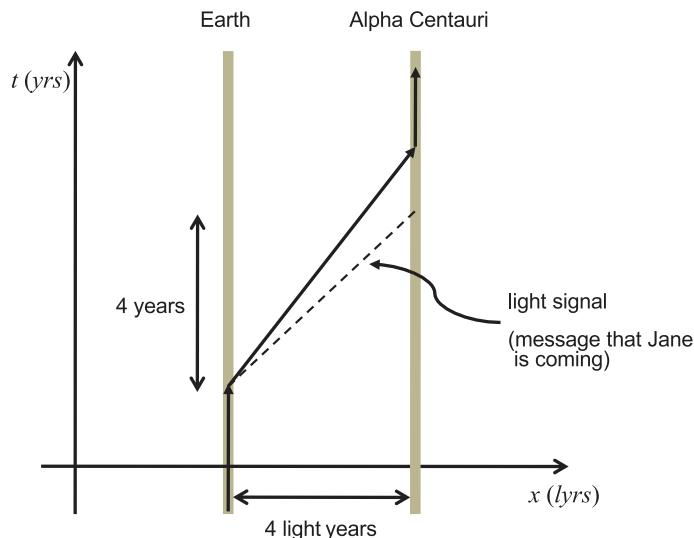


Fig. 3.18 Spacetime diagram, showing time on the *vertical axis* measured in years and distance on the *horizontal axis* measured in light years. Light signals are represented by 45° lines in these units. Also shown is the worldline of the Earth, Alpha Centauri, and the worldline (*thin black line*) of Jane's hypothetical trip to our nearest stellar neighborhood. Notice that the part of Jane's worldline representing her trip is at a steeper angle than 45° to the *horizontal axis*. This makes sense, because massive objects have to travel slower than the speed of light

Also shown in Fig. 3.18 is the worldline of a hypothetical trip Jane takes to Alpha Centauri, plus the worldlines of her friends who stayed on Earth and the hosts at Alpha Centauri who are waiting to receive her.

3.9 Causality in Spacetime

With space and time intervals varying from one observer to the next, you may feel that physical reality is slipping away. Is everything relative in Einstein's world, or is there some objective reality that we can hold on to? Especially worrisome is the relation between cause and effect. Is it possible, for example, that some observers will see Jane arrive at Alpha Centauri *before* she leaves Earth? It is reassuring that special relativity does not allow such bizarre occurrences.

Imagine an explosion that takes place at some time and location. Let's call this event A, as shown in Fig. 3.19. Light will propagate outwards along the 45° lines. Since physical influences cannot propagate faster than

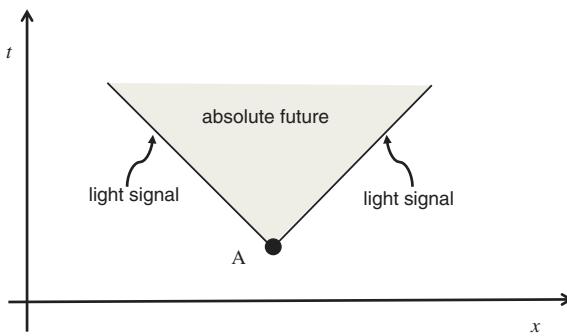


Fig. 3.19 A spacetime diagram showing event A, and the absolute future of A

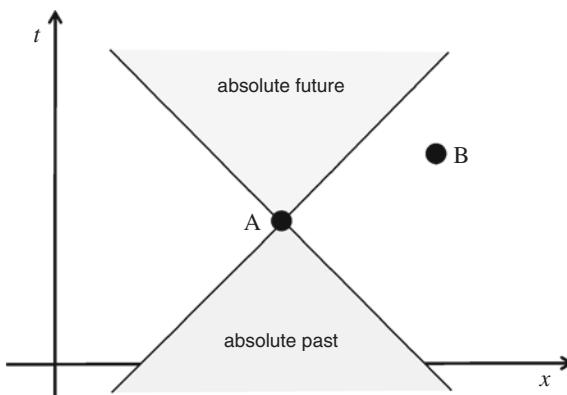


Fig. 3.20 A spacetime diagram showing event A, and the absolute past and future of A

light, matter particles from the blast will travel at speeds less than that of light, and thus if we could trace the worldlines of such particles, we would find that they are confined to lie in the shaded region in the figure. Simple mathematics shows that all observers will agree that all the events in this region are in the future of A (see question 12 at the end of this chapter). That is why this region is called the *absolute future* of event A.

We can also ask, which events can influence event A? In Fig. 3.20, events lying in the region denoted “absolute past” can have an influence over event A. Also shown is event B which can neither influence, nor be influenced by event A. For events in the absolute past of A, all observers will agree that they occurred earlier than A. However, for events like B that are not in the absolute future or past of A, observers will disagree: some of them will find that B is prior to A; others will find that it is later; and there will be some who will find that both events occurred at the same time.

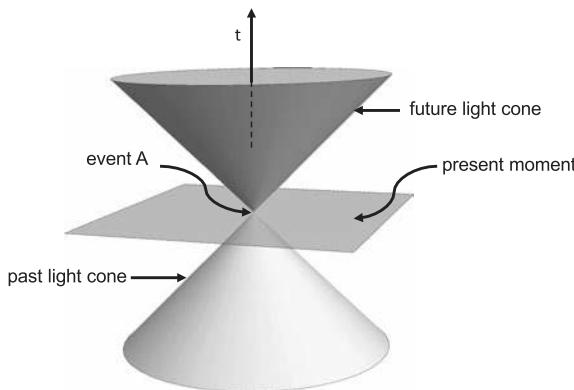


Fig. 3.21 3 dimensional spacetime diagram showing *light cones*. The two dimensional *slice* of the present is also shown

It follows that the time ordering of events can be reversed by going to another frame of reference only if the events are not causally related. If event B is caused by event A, all observers will agree that A occurred prior to B.

In the 2 dimensional spacetime diagrams of Figs. 3.19 and 3.20, light rays travel along 45° lines. If we add in another spatial dimension, so that we have a 3 dimensional spacetime diagram (as shown in Fig. 3.21), then these 45° lines are replaced by conical surfaces, called light cones. All events occurring within the *future light cone* (top), lie in the future of event A, and all events lying within the *past light cone* (bottom), occurred in its past.

Einstein completed his special theory of relativity in less than six weeks of frenzied work. Starting with two simple postulates, he was led by relentless logic to demolish the Newtonian concepts of absolute space and time, revealing a universe where observers in relative motion disagree on even the most basic measurements of mass, length and time. These counter-intuitive differences disappear when the observers' relative speeds are slow compared to light. In this limit, Einstein's theory reduces to that of Newtonian physics. In other words, Einstein did not prove that Newton was wrong. He just showed that Newton's theory has a limited range of validity, and that it is superseded by special relativity when relative speeds get close to the speed of light.

The word “special” in special relativity refers to the fact that this theory applies only in special circumstances when the effects of gravity are unimportant. This limitation is removed in Einstein's general theory of relativity, which is essentially a theory of gravitation.

Deriving the time dilation formula

Let's now derive Eq. (3.2). As before, consider an observer equipped with a clock consisting of a light source and a mirror at a distance L (see Fig. 3.11). When the clock is at rest, a light pulse will take time

$$t_0 = 2L/c \quad (3.7)$$

to complete the round trip from the source to the mirror and back again.

Now consider what happens when the clock is moving relative to the stationary observer with velocity v . The observer will see the light pulse travel along a diagonal path (see Fig. 3.11). The length of this diagonal path, $2D$, can be found using the Pythagorean theorem,⁷ thus

$$D = \sqrt{\left(\frac{vt}{2}\right)^2 + L^2} = \sqrt{\left(\frac{vt}{2}\right)^2 + \left(\frac{ct_0}{2}\right)^2} \quad (3.8)$$

where we used Eq. (3.7) to express L in terms of t_0 , and we define t to be the time for the light pulse to complete a round trip, as measured by the stationary observer. Thus, $t/2$ is the time it takes the light pulse to travel from the source to the mirror.

Because the observer must measure the speed of light to be c , he will measure the time taken for the round trip to be

$$t = 2D/c \quad (3.9)$$

Thus substituting $D = tc/2$ in Eq. (3.8), and squaring both sides we find

$$\frac{t^2 c^2}{4} = \left(v^2 t^2 + c^2 t_0^2\right)/4 \quad (3.10)$$

Rearranging, we find

$$t^2 = \frac{t_0^2}{(1 - v^2/c^2)} \quad (3.11)$$

which, upon taking the square root, gives the time dilation formula Eq. (3.2), with the Lorentz contraction factor defined as in Eq. (3.3).

Summary

Special relativity is based on two postulates: (1) the laws of physics are the same for all inertial observers (this is the *principle of relativity*), and (2) the speed of light in a vacuum is the same for all inertial observers. The constancy of the speed of light led Einstein to deduce that space and time intervals are relative: they depend on the state of motion of the observer who

⁷Recall for a right-angle triangle $h^2 = a^2 + b^2$, where h is the length of the hypotenuse, and a and b are the lengths of the other two sides.

measures them. In particular, he showed that the simultaneity of events is observer-dependent; moving clocks run slower than clocks at rest (time dilation); and moving meter sticks are contracted in the direction of motion (length contraction). He also showed that mass and energy are equivalent, $E = mc^2$.

Questions

1. The principle of relativity says that the laws of physics are the same for (choose one):
 - (a) all observers
 - (b) all observers moving in a straight line
 - (c) all observers moving in a straight line at a constant speed relative to an inertial reference frame.
2. You are in a room that is slowly rotating about a vertical axis. Are you an inertial observer? Can you do any experiment to detect the rotation of the room?
3. If you were riding on a train moving at constant speed along a straight track and you dropped a ball directly over a white dot on the floor, where would the ball land relative to the dot?
4. What distinguishes different parts of the electromagnetic spectrum? What do ultraviolet and infrared light have in common?
5. What are the two key postulates of Einstein's theory of special relativity? In what way does one of the postulates follow from the other?
6. Is it possible for your mother to go on a space trip and return younger than you? Is it possible for her to return younger than she was when she left?
7. Your twin's spaceship moves at 0.995 of the speed of light. If she left Earth on your birthday and travelled until she celebrated another birthday on the ship before returning home, who would be older when she returns, and by how many years? (Assume that your twin starts to head home at the same speed immediately after her birthday.)
8. A space traveler makes a round trip from Earth to a distant planet, moving at 0.8 of the speed of light, and comes back in 30 years as measured by his clock. How far away is the planet?
9. In order to accurately determine the location of your car, the clock on the GPS satellite has to be synchronized with the clock on Earth to an accuracy of 3×10^{-8} s. The satellite rotates about the Earth at the speed of 4 km/s; as a result time on the satellite runs slower by the factor

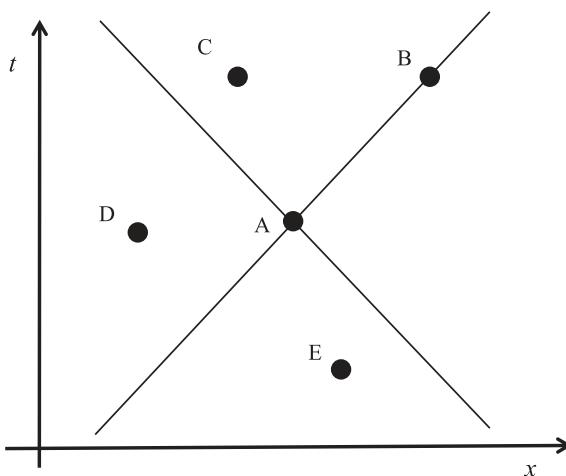


Fig. 3.22 Causality of space time events

$\sqrt{1 - v^2/c^2} \approx 1 - 9 \times 10^{-11}$. This factor is very close to 1, but the difference accumulates over time. What would be the discrepancy between the clocks on the satellite and on Earth after one day of operation⁸? Would GPS work if the effect of time dilation were not taken into account in the clock design? Why do we say that clocks run slower on a satellite than on Earth and not vice versa?

10. How fast is your alien friend's spaceship zipping past the Earth if we observe it contracted by 50%? By 99%? Is it possible for the spaceship to travel fast enough that it contracts to size zero?
11. In our day to day experience, why are we generally unaware of special relativistic effects?
12. In this spacetime diagram, which events can influence event A? Which events can A influence? Can event D influence any other events? (Fig. 3.22)
13. Consider the following spacetime diagram. Which worldlines (if any) represent the motion of: (i) an inertial observer? (ii) an accelerated observer? (iii) an inertial observer at rest with respect to the coordinate frame shown here? (Fig. 3.23)

⁸In this problem we are only concerned with the special relativistic effect of time dilation, which makes the satellite clock run slower. But there is also a general relativistic effect which makes time run faster for the satellite, because it is in a weaker gravitational field than a clock on Earth (we will discuss this in Chap. 4). If we were to take into account both effects, then we would actually find that the satellite clock runs faster overall.

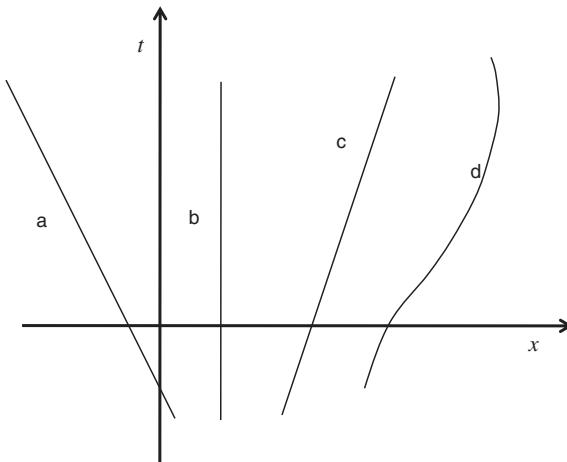


Fig. 3.23 Spacetime diagram with several worldlines

14. What is the difference between invariant and relative quantities? Can you describe two quantities that are invariant in Newtonian mechanics, but are relative in special relativity? Can you name two quantities that are invariant in special relativity?
15. Suppose you travel to an interstellar vacation spot, which is 4 light years away. Also suppose you arrive there 5 years later as measured by the clocks of your friends on Earth.
 - (a) Find the spacetime interval (in light years) between your departure and arrival.
 - (b) Use the invariance of the spacetime interval to find the time elapsed by your own clock between your departure and arrival. (Hint: in your own reference frame, you did not move, so the space separation is $\Delta x = 0$.)
16. Use the invariance of the spacetime interval to show that if event B is in the absolute future of event A, then all observers will agree that B occurred later than A.
17. An interstellar merchant is contemplating the following transaction. He buys some goods here on Earth, transports them to Barnard's star (about 6 light years away), exchanges them for Barnardian goods, flies them back to Earth and sells them here for profit. Normally, he would regard this transaction profitable if it brings more profit than the same capital investment would bring in the same time at the going interest rate. But now he is not sure what time he should use in this calculation: should it be the time elapsed on Earth or on the spaceship making the round trip

to Barnard's star? These times are rather different, since the spaceship is going to travel at a speed close to that of light! What do you think? Does it matter whether the merchant himself takes the trip or stays on Earth? (This and related issues are discussed in the tongue-in-cheek article “The Theory of Interstellar Trade” by the Nobel Prize winning economist Paul Krugman, published in the March 2010 issue of the journal *Economic Inquiry*. Krugman noted that his article “is a serious analysis of a ridiculous subject, which is of course the opposite of what is usual in economics.”)

4

The Fabric of Space and Time

Einstein's special theory of relativity was a great breakthrough in our understanding of the physical world, but it presented a problem: it was incompatible with Newton's law of gravitation. Newton himself, and some ten generations of physicists and astronomers that followed, used this law to describe the motion of planets with remarkable accuracy. Granted, there was a tiny discrepancy in the calculated precession rate of the orbit of Mercury, but it did not seem to be a cause for concern. After all, theories seldom agree with all the data at any given time. Some of the data may simply be wrong and some discrepancies are later explained away with more careful theoretical analysis. Thus, Newton's theory appeared rock solid. However, it did not fit into the framework of special relativity.

The inconsistency between the two theories is not difficult to illustrate. Newton's theory states that the force between two bodies is inversely proportional to the square of the distance between the bodies at an *instant of time*. But according to Einstein, the distance (and the notion of an “instant of time”) is not the same for different observers. So whose distance should we use? If Newton's law is valid in the reference frame of some preferred inertial observer, while it is not valid for other inertial observers, then the principle of relativity is violated, because it requires that physical laws should apply equally to all inertial observers. So clearly, Newton's theory or relativity had to go ...

4.1 The Astonishing Hypothesis

In his *Dialogue*, Galileo argued that the motion of objects under the action of gravity is independent of their mass, size, or any other intrinsic properties, as long as air resistance and other non-gravitational forces can be neglected. This was contrary to the accepted Aristotelian viewpoint at the time, which claimed that heavier objects fall faster. Indeed, a cannon ball does fall faster than a feather, but Galileo realized that the difference was due only to air resistance. Legend has it that Galileo dropped rocks of different mass from the Leaning Tower of Pisa, to see if they hit the ground at the same time. We do know that he experimented with marbles rolling down inclined planes and found that the motion was independent of the mass. He also offered the following theoretical proof that Aristotle could not be right: suppose that indeed a heavy rock falls faster than a light rock. Imagine then tying them together with a very light string. How will this affect the fall of the heavy rock? On the one hand, the slower-moving light rock should make the fall of the heavy rock somewhat slower than it was before. On the other hand, viewed together the two rocks now constitute one object, which is more massive than the heavy rock was initially, and thus it should fall faster. This contradiction demonstrates that Aristotle's theory is inconsistent (Fig. 4.1).

Galileo's experiments and theoretical deductions thus revealed that all objects fall at the same rate, regardless of their mass. While Einstein was pondering this peculiar kind of motion, which is completely independent of the object that is moving, it reminded him of inertial motion. Remember



Fig. 4.1 Galileo's thought experiment

that in the absence of forces, an object moves at a constant velocity along a straight line in spacetime, *regardless of what the object is made of*. It is as if the path of the object in space and time is a property of spacetime itself.

The analogy between motion in the field of gravity and inertial motion goes even further. Suppose that instead of the cabin of a ship, as suggested by Galileo, you lock yourself up in a falling elevator. (This is recommended, of course, only as a thought experiment!) All objects in the elevator, and the elevator itself, will fall at the same rate. You will not feel your weight because the elevator floor will be falling under your feet. If you drop an object, it will float next to you, exactly as it would if you were an inertial observer far away from any gravitating bodies. The same state of weightlessness is experienced by astronauts when their spaceship moves in the gravitational field of the Earth with its engines turned off. Indeed, the motion of objects in the gravitational field does look very similar to inertial motion. But there is also a difference—gravity makes objects accelerate towards the Earth's center, so their worldlines are no longer straight ... (Fig. 4.2).

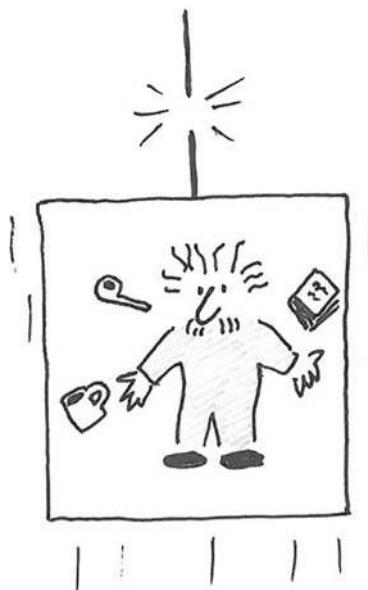


Fig. 4.2 Weightlessness inside a falling elevator

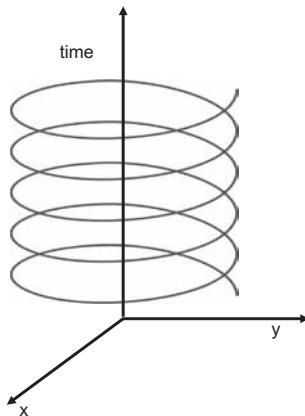


Fig. 4.3 Earth's worldline as it revolves around the sun

This line of thought led Einstein to an astonishing hypothesis: in the presence of gravity, objects still move along the straightest lines in spacetime, *but the spacetime itself is curved*. The idea is that massive bodies curve spacetime around them. For example, the spacetime in the vicinity of the Sun gets curved. Thus, the Earth does not move along a straight line at a constant velocity (as it would in the absence of any gravitating bodies), but instead it moves around the Sun. The Earth's worldline is in fact the straightest line in this curved spacetime. (Such lines are also called geodesics). Note that a geodesic line in *spacetime* does not necessarily correspond to the straightest trajectory in *space*. For example, the Earth's elliptic orbit around the Sun is certainly not the straightest possible path (Fig. 4.3).

The distortion of spacetime geometry by a massive body is often illustrated by a heavy bowling ball resting on a horizontally stretched rubber sheet. The rubber surface is warped near the ball, just like the spacetime is warped near a gravitating body. If you roll a marble along the rubber surface, its path will be curved, due to the warping of the sheet. The trajectory of the marble is analogous to that of light signals and small objects. Note, however, that the time dimension is suppressed in this picture, so it illustrates only the warping of space and not of spacetime (Fig. 4.4).

A curved four-dimensional spacetime is an abstract concept; it is very difficult to visualize. We shall now take an excursion to develop some intuition about spacetime curvature, using lower-dimensional analogies. As a first step, we shall leave time aside and address a simpler issue: What does it mean for space to be curved?

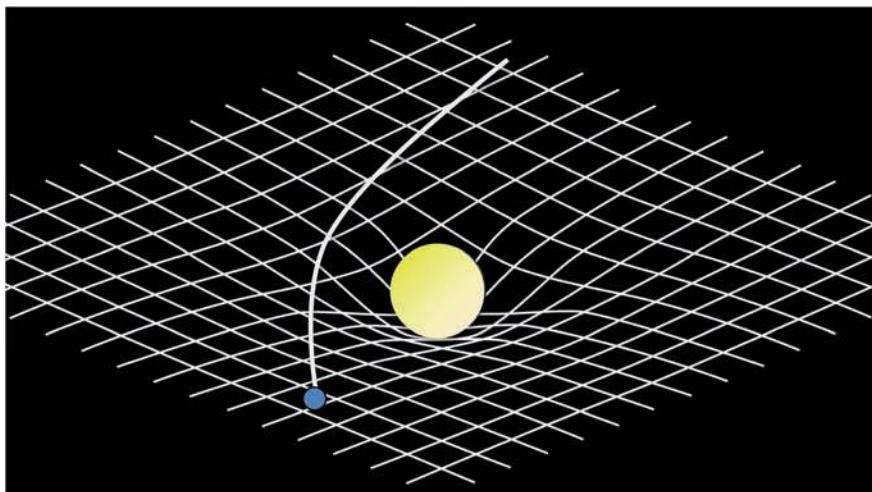


Fig. 4.4 Spacetime curves around a massive body, just like a stretched rubber sheet warps when a bowling ball is placed on it

4.2 The Geometry of Space

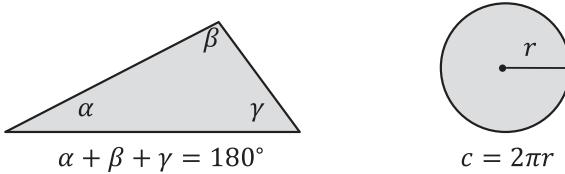
4.2.1 Euclidean Geometry

The geometers of ancient Greece devoted a great deal of effort to studying the properties of space. A beautiful exposition of their work was compiled around 300 B.C. by Euclid of Alexandria in his book *Elements*, which is regarded as the most influential text in the history of mathematics. Euclid began with five *axioms*. These are self-evident statements whose truth would be indisputable by any sane person. For example, the first two axioms are: “Given any two points, a straight line segment can be drawn between them”, and “Any straight line segment can be extended without limit in either direction.” If you accept Euclid’s five axioms, you have no choice but to follow the inevitable logic of the proofs of 465 theorems which express various geometrical facts, including the following (Fig. 4.5):

The sum of the angles in any triangle is 180° .

The circumference of a circle of radius r is $C = 2\pi r$.

The area of a sphere of radius r is $A = 4\pi r^2$.

**Fig. 4.5** Euclidean geometry

The volume of a sphere of radius r is $V = \frac{4}{3}\pi r^3$.

This amazing creation of the Greeks was so perfect that mathematicians remained under its spell for more than 2000 years. Euclid's axioms looked as obvious and necessary as the laws of logic themselves. Thus it appeared as though the properties of space could be deduced from pure reason. Furthermore, Euclidean geometry seemed to be the only geometry that was logically possible (Fig. 4.6).

Occasional doubts had only been expressed regarding Euclid's fifth axiom, which states: "Given any straight line, you can draw one and only one straight line parallel to it through any point in the same plane." (It is assumed that the point is not on the first line.) Looking at Fig. 4.7, this statement seems rather plausible, although it is perhaps not as obvious as Euclid's other axioms. Numerous attempts were made to prove it as a theorem (this would reduce the number of axioms to four). However, as the centuries passed, no one raised the possibility that the parallel line axiom might actually be wrong, or that it might be possible to replace it with something else that is free from logical contradictions.

Our intuition is rooted in Euclidean geometry, which very accurately represents the properties of space—at least on the scales familiar to humans. However, imagine for a moment that you want to test Euclid's fifth axiom experimentally. First you have to decide what you mean by a straight line. Of course you can draw a line with a ruler. But how do you know that your ruler is straight? You can check it using a stretched thread, or by holding it close to your eye and looking along its length. But then you are assuming that the thread is straight, or that light propagates in a straight line. Clearly, you have to choose some class of objects and identify them with straight lines; otherwise you have no standard of straightness.

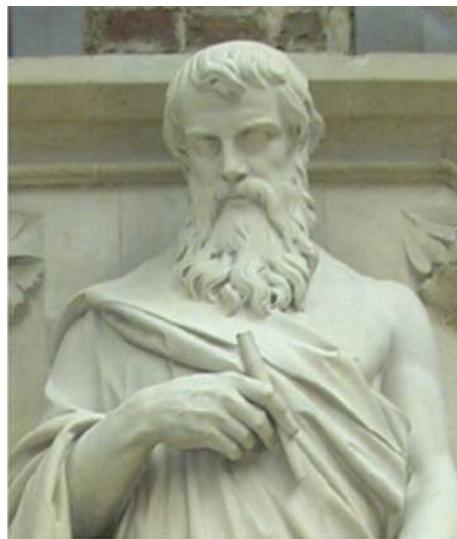


Fig. 4.6 Euclid of Alexandria. Credit Statue of Euclid Oxford University Museum of natural history

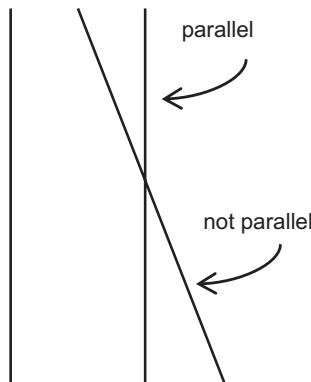


Fig. 4.7 Parallel lines in Euclidean geometry. By definition, straight lines are called parallel if they do not intersect

For the sake of argument, suppose you choose light rays to be your standard straight lines. Then imagine shining two beams of light from two projectors onto a distant screen. You make sure that the beams emanate in the same direction, orthogonal to the line connecting the projectors. If Euclid

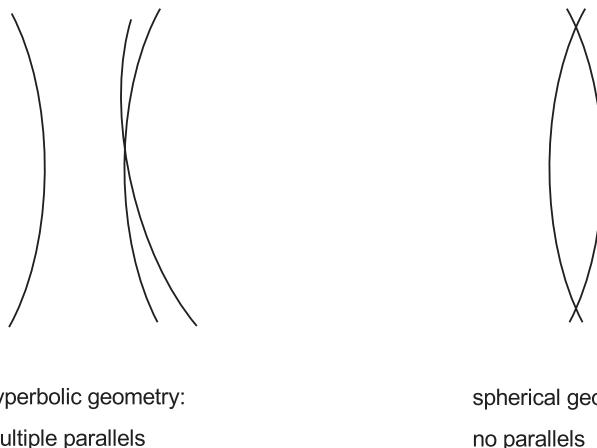


Fig. 4.8 *Multiple parallel lines and no parallel lines in non-Euclidean geometry*

is right, then the distance between the two light spots on the screen should be equal to the distance between the projectors. But imagine that the light spots turn out to be slightly further apart, and that this distance keeps increasing as you move the screen further away. This would mean that the separation between the light beams grows with the distance. If your straight lines have this property of somehow “bending away” from one another, then it is not difficult to imagine that they can avoid intersection even if initially they are headed slightly towards each other. A number of different lines could then pass through the same point without ever intersecting a given line (see Fig. 4.8). Alternatively, if the lines were to bend towards one another, then they might always cross. This is also illustrated in Fig. 4.8. If light rays in our world had such properties, then perhaps we would not find Euclid’s fifth axiom so obvious.

4.2.2 Non-Euclidean Geometry

The great German mathematician Carl Friedrich Gauss was the first to break away from the Euclidean dogma. In the early 1800s he explored a geometry in which the fifth axiom is replaced with a postulate allowing multiple parallel lines through the same point, and convinced himself that it was free from logical contradictions. He worked out the properties of various geometrical figures and found that they were in many ways different from those in the familiar Euclidean geometry. In particular, the sum of the angles in a triangle was always less than 180° . This kind of geometry is now called *hyperbolic* geometry.

With a breakthrough of this magnitude, you might imagine that Gauss ran around the streets of Gottingen shouting “Eureka!”, and then immediately submitted a paper for publication. But this did not happen—Gauss kept his work secret! In his time, challenging Euclid was what would now be called “politically incorrect”. Euclidean geometry was adopted without question by the great Newton, and was declared “an inevitable necessity of thought” by the eminent German philosopher Immanuel Kant. It was not uncommon for academics to get embroiled in life-long bitter disputes, and Gauss probably felt that passing on a publication was a reasonable price to pay to avoid any altercations.

Hyperbolic geometry was independently discovered in the 1820s by Nikolai Lobachevsky, a professor of mathematics in the provincial Russian city of Kazan, and a few years later by a Hungarian artillery officer Janos Bolyai. Lobachevsky was not afraid to stick his neck out and submitted his work for publication to the St. Petersburg Academy of Sciences. His paper, however, was rejected and was finally published in the obscure *Kazan Messenger*. Lobachevsky’s discovery did not receive much recognition during his lifetime, and at the age of 54 he was suddenly dismissed from his post at the University. No reason was given, but this might well have something to do with his unorthodox ideas. The hyperbolic geometry is now often called Lobachevsky geometry.

The geometry in which no lines can be drawn parallel to a given line through any point, was investigated in the 1850s by Bernhard Riemann, who would later become Gauss’s successor as a professor at Gottingen. It is sometimes called spherical geometry, for reasons that will soon become clear.

4.3 Curved Space

Apart from his secret research on hyperbolic geometry, Gauss developed another, completely different approach to the problem. This work was even more important, because of its greater generality. For us it also has an added benefit of being very useful for visualizing non-Euclidean spaces.

4.3.1 The Curvature of Surfaces

For nearly a decade Gauss was involved in large-scale geodetic measurements. The effort was well funded, in order to produce accurate maps, but Gauss’s own interest was to gain more information about the shape of the



Fig. 4.9 Carl Friedrich Gauss (1777–1855) was a child prodigy born to poor parents. A caring teacher managed to convince Gauss's father to excuse the young Gauss from his part-time job spinning flax, so that he may continue his education. Word of his talents reached the Duke of Brunswick, who sent him to grammar school, and then to university at the age of 15. Gauss made major discoveries in many areas of mathematics and physics and was already in his lifetime regarded as one of the greatest mathematicians who ever lived. His fame led to Napoleon giving a command to spare Brunswick because "the foremost mathematician of all time lives there"

Earth. The shape of a surface can be easily grasped when you see it from outside. But satellite pictures of the Earth were not yet available, and Gauss could rely only on measurements made on Earth's surface. He was thus led to think about the inner properties of a surface (Fig. 4.9).

Gauss's key realization was that a surface could be regarded as a two-dimensional space in itself, as if the exterior world were non-existent. The role of a straight line connecting two points in that space is played by the geodesic line, which is the shortest line between two points along the sur-

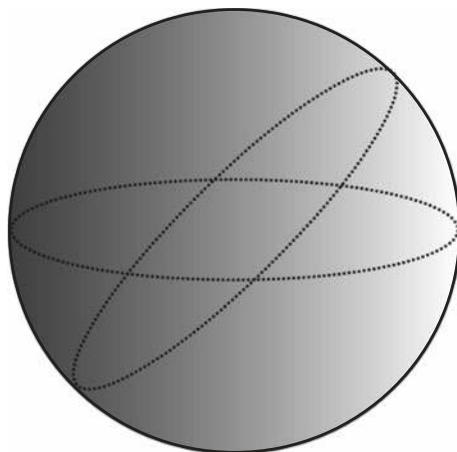


Fig. 4.10 Geodesic lines are great *circles* on a *sphere*

face. While a triangle on a flat surface has angles which sum to 180° , this is not the case for a triangle made of three geodesic lines on a non-flat surface. Gauss found that for small triangles the deviation from 180° grows with the area of the triangle and is proportional to a quantity that he called the curvature of the 2D space. (From now on we shall use the abbreviation 2D to denote “two-dimensional”.) The curvature can be positive or negative, depending on whether the sum of the angles is greater or smaller than 180° (Fig. 4.10).

A simple prototype of a curved surface is a sphere. The geodesic lines in this case are great circles. For example, meridians and the equatorial circle are geodesics on the globe. Any two great circles necessarily intersect—like all meridians intersect at the North and South poles, even though they appear to be parallel near the equator. This means that parallel lines do not exist on a sphere, and thus Euclid’s fifth axiom does not hold. Also, triangles that are constructed from great circle segments, have angles which add to more than 180° (see Fig. 4.11b). Gauss found that the curvature of a sphere is inversely proportional to the square of its radius. As the radius is increased, the curvature gets smaller, and in the limit of infinite radius the curvature vanishes and the inner geometry is Euclidean. In this limit the sphere is indistinguishable from a plane.

A 2D surface of negative curvature can be pictured as a saddle-like surface, as in Fig. 4.11c. In general, some parts of a curved surface may have positive and other parts negative curvature. The case of a sphere is rather

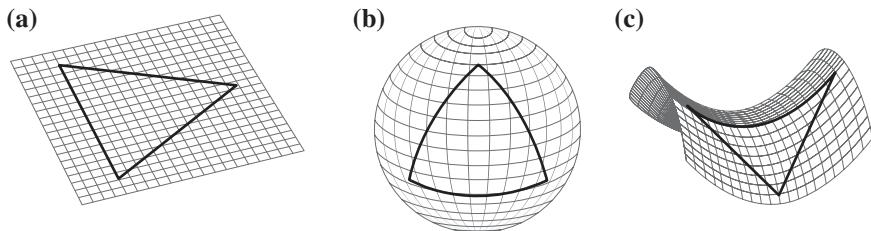


Fig. 4.11 Two dimensional spaces **a** flat Euclidean space; **b** spherical space; **c** hyperbolic space. As noted in the text, we cannot actually draw a 2D hyperbolic space in our 3D Euclidean space; the saddle is a space of negative curvature, which is often used as a “stand in” for a hyperbolic surface (Springer Artist)

special, because of its high symmetry. It is *homogeneous*, which means that no matter where you are on the sphere, it looks the same. It is also *isotropic*, which means that from any point on a sphere it looks the same in all directions. A homogeneous and isotropic 2D space of negative curvature is called a hyperboloid, and the inner geometry of that space is the hyperbolic geometry of Gauss and Lobachevsky. We would be glad to include a photograph of a hyperboloid in the book, but unfortunately this surface cannot be embedded in a 3D Euclidean space. Gauss’s work on the inner geometry of surfaces was later extended by Riemann to spaces of three and higher dimensions.

4.3.2 The Curvature of Three-Dimensional Space

Of special interest are the homogeneous and isotropic 3D spaces in which all locations are equivalent, and which look the same in all directions. As in the case of two dimensions, there are only three types of such spaces: *Euclidean*, *spherical* and *hyperbolic*. Euclidean (or flat) space is the usual space we are most familiar with. It has zero curvature, and an infinite volume—it goes on and on ad infinitum.

Spherical space (or the 3D sphere) is a three-dimensional analogue of a 2D spherical surface. It is a closed, finite space; its geodesics are circles of length $2\pi R$, and its volume¹ is $2\pi^2 R^3$. The parameter R is called the radius of the 3D sphere. The sum of angles in a triangle in this space is greater than 180° . Also, as the radius of curvature R gets very large, the 3D curved space approaches 3D Euclidean space.

Our everyday experience suggests that we live in a 3D Euclidean space. But what if we really live in a 3D spherical space with an astronomically large radius of curvature? Would we be able to tell the difference?

There is an observational way to make the distinction that hinges on an unusual property of spherical space. Let us consider how the area of a 2D sphere, $A(r)$, depends on its radius in this space. Since we cannot visualize a curved 3D space, we shall use a 2D analogy. Imagine lines of latitude on the Earth's surface. We can think of them as 1D "spheres" centered at the North Pole. The distance r , along the Earth's surface, from the North Pole to a given line of latitude plays the role of radius of the 1D "sphere" (see Fig. 4.12). As the radius is increased, the circumference of the latitude lines grows from zero up to the maximum value, $c_{max} = 2\pi R$, where R is the radius of the Earth. But as we further increase the radius past the equator, the circumference starts to diminish, coming back to zero at the South Pole. By analogy, we expect the area of a 2D sphere, in a spherical space of radius R , to increase from zero to $A_{max} = 4\pi R^2$ and then decrease back to zero.

The dependence of the area on the radius, $A(r)$, is important because it determines how the observed brightness of a light source depends on the distance of the source from the observer. The energy emitted by the source per unit time is uniformly distributed over the area of the sphere. Hence, the observed brightness of the source is inversely proportional to $A(r)$. In flat space, $A(r) = 4\pi r^2$, and the brightness decreases as r^{-2} with the distance (an inverse square law). On the other hand, in a spherical space the brightness initially decreases with distance, but at $r > \frac{\pi}{2}R$ it starts growing again, and becomes very large as r approaches the maximal distance, $r_{max} = \pi R$ (the "South Pole"). If we think of meridians in Fig. 4.12 as light rays emanating from a source at the North Pole, these rays start to converge after crossing the equator and focus to a point at the South Pole. An observer near the South Pole will therefore see a very bright image of the source.

A 3D hyperbolic space is a space of constant negative curvature. Its volume is infinite, and the sum of angles in a triangle is less than 180° . The area of a sphere grows faster than r^2 , so light sources get dimmer with distance even faster than they do in flat space. The 3D hyperbolic geometry is even harder to conceptualize than spherical curved geometry, so we will not go into any more detail here. Fortunately, like Euclidean geometry, hyperbolic and spherical geometries can be readily described mathematically.

When curved non-Euclidean spaces were first shown to be logically consistent, Gauss attempted what appeared to be an observational test of curvature. Using an instrument that he himself invented for geodetic

¹The *volume* of a 3D spherical space is analogous to the *area* of a regular 2D sphere. Notice that this volume is different from the volume enclosed by a sphere in 3D Euclidean space $V_E = \frac{4}{3}\pi r^3$.

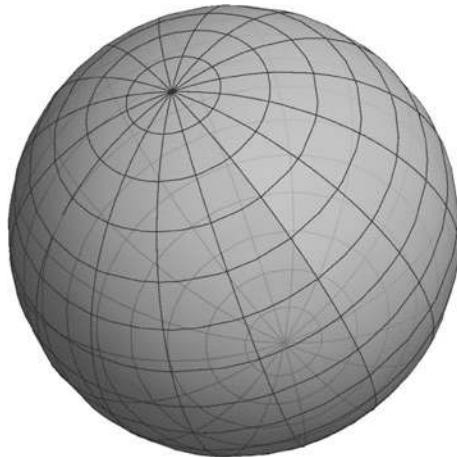


Fig. 4.12 Lines of latitude can be thought of as “1D spheres” in a 2D spherical space. The distance from the North Pole plays the role of radius of the “spheres”. The circumference of the “spheres” grows with the radius, reaches its maximum at the equator, and then decreases, shrinking to zero at the South Pole

measurements, the heliotrope, Gauss measured the three angles in a triangle having a size of about 100 km, with vertices at the mountain tops of Hohenhagen, Brocken and Inselsberg. He mentioned this measurement in a paper published in 1827, saying that the angles added up to 180° within the expected errors of measurement. It is left for the reader to decide whether this was only a test of the accuracy of the heliotrope, or a test of the Euclidean geometry of space. If it was the latter, Gauss’s attempt came too early: the curvature of space had to await its detection for nearly 100 years.

4.4 The General Theory of Relativity

After this foray into non-Euclidean geometry, we now return to Einstein and his struggle to understand gravity. It took Einstein five years to leap from Galileo’s cue to the idea of curved spacetime. This was a tremendous breakthrough, but it was not nearly the end of the journey. The idea still had to be expressed in mathematical terms: precisely how is the curvature of spacetime determined by massive bodies? For that matter, how can one mathematically characterize a curved four-dimensional spacetime? Einstein had no idea.

He approached his old classmate Marcel Grossmann, a mathematician and an expert in geometry. After some time in the library, Grossmann reported back with some good news and some bad news. The good news was that the mathematics of curved spaces did exist. It had been developed by Bernhard Riemann, who extended Gauss's work on the non-Euclidean geometry of curved surfaces to spaces of three and higher dimensions. The bad news was that this mathematics was an impenetrable mess. In the case of a surface, the curvature is characterized by a single number (at every point), while in higher dimensions it is described by a multi-component monster called the Riemann tensor. Physicists would have been well advised to stay away from this...

But Einstein did not have this option. With Grossmann's help, he mastered the intimidating formalism of Riemannian geometry and went on to use it in the formulation of his new theory of gravity. "... in all my life", he wrote in a letter to German physicist Arnold Sommerfeld, "I have never struggled so hard ... Compared to this problem, the original relativity theory is child's play." It took Einstein more than three years to complete the job.

The equations of the new theory, which Einstein called the "general theory of relativity", relate the geometry of spacetime to the material content of the Universe, see Fig. 4.13. This may look like a single equation, but the indices μ and ν take four possible values: 0, 1, 2, 3, so in fact this is a compact representation of a system of 16 equations. The left-hand side of the equations includes components of the Riemann tensor, which tell us

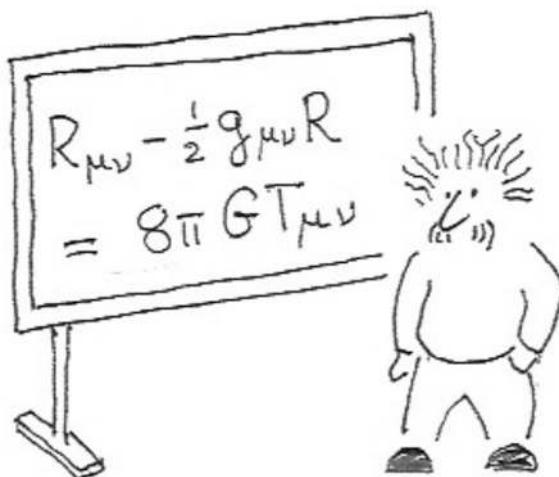


Fig. 4.13 Einstein's equations

how much the spacetime is curved in different directions. The right-hand side contains Newton's constant G and the so-called stress-energy tensor of matter $T_{\mu\nu}$, whose components include the energy density, the energy flux (which characterizes how fast energy is being transported and in what direction), and pressure.

You may be surprised to see all these different characteristics of matter, because in the Newtonian theory the gravitational force is determined only by mass. But the difference here is not as large as it may seem. The energy density is essentially the same thing as mass density (remember $E = Mc^2$?), while the other quantities have little effect on gravity under normal circumstances. The energy flux is small when the speeds of gravitating bodies are well below the speed of light, and the pressure is normally much smaller than the energy density. Later in this book we shall encounter exotic states of matter with a very large pressure, but for familiar astrophysical objects like stars or planets, the role of pressure in gravitational physics is negligible.

An essential feature of Einstein's theory is that gravitational effects propagate at the speed of light. If the Sun were suddenly removed, as if walloped by a huge golf club, this would first change the spacetime curvature only in its immediate vicinity. The effect would then spread out and reach the Earth in about 8 min (the time it takes light to travel from the Sun to the Earth). Thus, the force of gravity is no longer determined by the instantaneous distance between the bodies. However, the Newtonian instantaneous interaction is a very good approximation when the motion of bodies is slow compared to the speed of light.

Einstein verified that in the Newtonian regime of slow motion and weak gravitational fields, his theory reproduced Newton's law, with the force of gravity inversely proportional to the square of the distance. In fact, he found that the difference between the two theories was completely negligible for the motion of planets in the Solar System. The only exception was Mercury, the planet closest to the Sun. Before general relativity, Mercury's orbit was measured to precess around the Sun by about 2° per century (see Fig. 4.14), instead of being perfectly elliptical. It was understood that the other planets perturbed Mercury's orbit, and the resulting precession rate could be calculated. However, the most detailed Newtonian calculations predicted a rate that was about 1% slower than observed. Einstein was aware of this discrepancy, and showed that a small correction to Newton's law due to general relativity precisely made up for the difference. Einstein needed no further proofs: at this point he was convinced that his theory was correct.

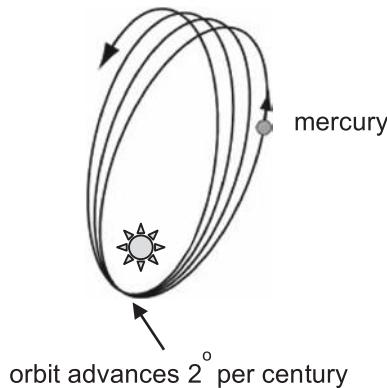


Fig. 4.14 The precession of Mercury's orbit (this figure is not to scale)

4.5 Predictions and Tests of General Relativity

4.5.1 Light Deflection and Gravitational Lensing

The first new prediction of general relativity to be tested observationally was the deflection of light as it propagates through a curved spacetime region near a massive body. If light from a distant star passes close to the Sun, the star should appear to be in a different position from where it usually is. Einstein proposed that a Solar eclipse would offer the perfect opportunity to view stars that appear close to the Sun. Their positions could then be measured and compared with their known positions when the Sun is not nearby.

In 1919 two British teams led by Arthur Eddington set out to test Einstein's prediction. Eddington's announcement to the world that they had indeed measured the bending of starlight in complete agreement with Einstein's theory, instantaneously turned Einstein into a household name. As for Einstein, he was so confident that when he was asked what if Eddington doesn't confirm the theory, he replied "Then I would feel sorry for the dear Lord!"

A related prediction is that of gravitational lensing: light from a distant source is bent as it passes by a massive object, like a galaxy, resulting in multiple images of the same source (see Fig. 4.15). If the lensing galaxy happens to be centered on the line of sight to the source, the images are spread into a circular band known as an Einstein ring, as shown in Fig. 4.16.

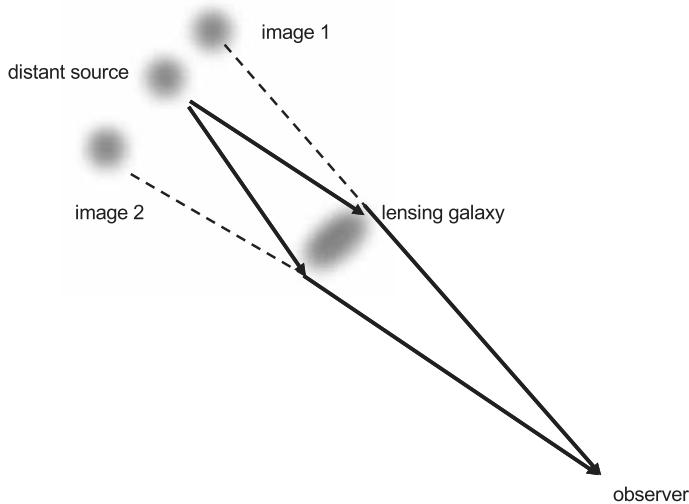


Fig. 4.15 Gravitational lensing

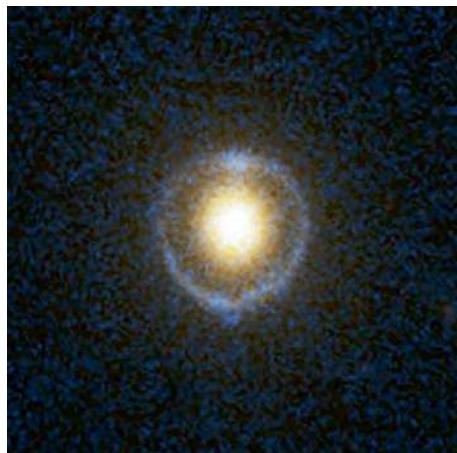


Fig. 4.16 Depending on the alignment of the observer, lens and source, it is possible for the image of the source to be spread into a *ring* around the lensing galaxy. Credit NASA , ESA , A. Bolton (Harvard-Smithsonian CfA) and the SLACS Team

4.5.2 Gravitational Time Dilation

Einstein also investigated how clocks are affected by gravity and found that a clock runs slower as it gets closer to a gravitating body. This effect is known as *gravitational time dilation*. Surprisingly, it plays an important role in our tech savvy lives. The GPS system in your car, and that used by airplanes too, depend on information coming from satellites in space. The positional information these satellites relay depends on the amount of time it takes light signals to travel to and from your device. But whose time? The satellites are moving much slower than the speed of light, but the effects of special relativity are non-negligible and cause the clocks on satellites to run a little slower—by about 7 μs per day. On the other hand, the satellites are in a weaker gravitational field than clocks on the surface of the earth, and thus by gravitational time dilation their clocks run a little faster—by about 45 μs per day, so the net result is that clocks on satellites run faster by about 38 μs a day. Both of these effects are accounted for in GPS satellite design; if they were not, position determinations would be so inaccurate that there would be no point to having a GPS!

4.5.3 Black Holes

General relativity predicts the existence of compact dense objects called black holes. Loosely speaking, the defining characteristic of a black hole is that a huge amount of mass is contained in a relatively tiny region of space. The spacetime in the vicinity of the mass is so severely warped that not even a light beam can find its way out of the region.

Let's quantify this idea. Consider an object of mass M and radius R . Recall from Chap. 2, the escape velocity from the surface of this object can be calculated from the condition that the total mechanical energy be zero: $E = \frac{1}{2}mv^2 - \frac{GMm}{R} = 0$, yielding

$$v_{esc}^2 = \frac{2GM}{R} \quad (4.1)$$

If the escape speed is equal to the speed of light, $v_{esc} = c$, then the object is a black hole. The radius at which this happens is the Schwarzschild radius,

$$R_s = \frac{2GM}{c^2}, \quad (4.2)$$

named after the German physicist Karl Schwarzschild, who found the solution of Einstein's equations describing a black hole.² For the Earth, the Schwarzschild radius is about 1 cm, which is why packing the Earth into the tip of your thumb would create a black hole! The Sun's Schwarzschild radius is 3 km [as you can check using Eq. (4.2)].

The spherical surface of radius R_s enclosing the black hole is called the *event horizon*. An observer outside a black hole can never see beyond this surface. To illustrate some unusual properties of the event horizon, imagine that your twin sister embarks on a daring space mission towards a black hole, while you stay at a safe distance outside. Her spaceship is equipped with a device that sends out a light pulse every second, according to the clock on board. You will notice that the light pulses get less and less energetic as the spaceship approaches the horizon. This is because the light pulse has to expend energy to climb out of the strong gravitational field near the black hole. At the event horizon the light will not be able to climb out at all.

Furthermore, you will notice that the time intervals between successive pulses get increasingly large. This is due to gravitational time dilation. When your sister gets very close to the event horizon her clock seems to stop, and she will appear frozen in time—at the event horizon. Thus you will never see her “cross” the event horizon—no matter how long you wait.

As far as your sister is concerned, she will not notice anything special when her spaceship approaches the horizon. While she is still outside, it is a good idea to turn around and head back. When she joins you, she will be younger than you are. If instead she stays the course, she will cross the horizon uneventfully, in a finite time by her clock. But that is the point of no return: the spaceship will now fall inexorably towards the center of the black hole. Close to the center, strong tidal forces stretch all falling objects in one direction and squeeze them in other directions, so the spaceship and its cargo will be “spaghettified”.

We will return to discussing black holes in Chap. 12, but note here that there is strong evidence that they do in fact exist.

4.5.4 Gravitational Waves

Another key prediction of general relativity is that accelerating massive bodies generate small distortions (or ripples) in the geometry of spacetime, called *gravitational waves*, that travel at the speed of light, much like accel-

²GR gives the same formula Eq. (4.2) for the Schwarzschild radius as our Newtonian derivation.

erated charges generate electromagnetic waves. Gravitational waves interact very weakly with matter and are therefore extremely hard to detect. Nonetheless, *indirect* evidence for gravitational waves has been known for some time. General relativity predicts that as two stars orbit one another, they release energy in the form of gravitational waves. This loss of energy causes the stars to spiral in towards one another, which increases their orbital speed, and decreases their orbital period. In 1974 a spiraling binary pair consisting of two neutron stars was detected. Over the last few decades, the orbital period of the pair has changed in the precise way predicted by general relativity.³

In September 2015 gravitational waves were *directly* detected using both of the Laser Interferometer Gravitational Wave Observatory (LIGO) detectors, located in Livingston, Louisiana, and Hanford, Washington, USA. The LIGO detectors were able to measure the distortion of spacetime caused by gravitational waves that emanated from a coalescing pair of black holes. Scientists deduced that 29 and 36 Solar mass black holes collided to form a larger spinning black hole. The collision occurred about 1.3 billion years ago, lasted for a fraction of a second, and converted about 3 Solar masses of energy into gravitational wave radiation, some of which passed through the Earth. The spatial distortion caused by gravitational waves is almost imperceptibly small: each LIGO instrument has two arms that are about 4 km long, and changes in its length of about one ten thousandth the size of a proton are measurable. It is a huge experimental triumph to be able to measure such a minute change in the apparatus, and to be able to deduce so much about the nature of the system which generated the gravitational radiation.

The detection of gravity waves has opened up an entirely new spectrum, akin to the electromagnetic spectrum, with which to observe the universe. Several other gravitational wave observatories are scheduled to become operational in the near future (Fig. 4.17).

Today, the scientific success of general relativity is unquestionable. But perhaps the most remarkable thing about GR is how little factual input it required. The postulate that Einstein used as the foundation of the theory, that the motion of objects under the action of gravity is independent of their mass, was already known to Galileo. With this minimal input, he created a theory that reproduced Newton's law in the appropriate limit and explained a deviation from this law. If you think about it, Newton's law is somewhat

³This discovery earned Joseph H. Taylor and Russell A. Hulse the 1993 Nobel Prize in physics.

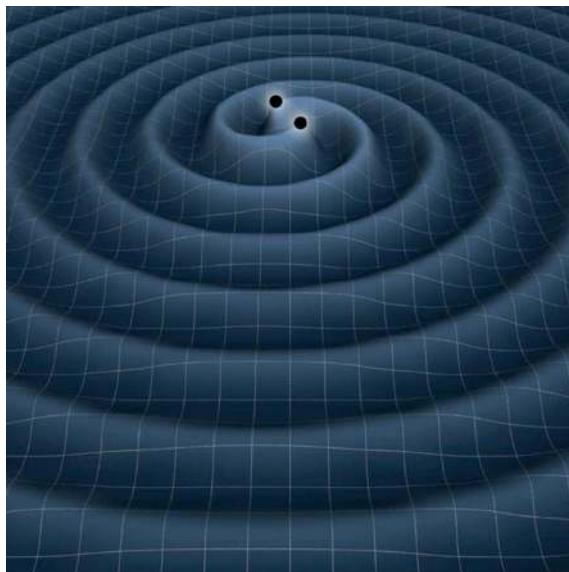


Fig. 4.17 Gravitational waves: illustration of two heavy orbiting masses generating ripples in spacetime. The amplitude of gravitational waves is expected to be much smaller than displayed here. Credit T. Carnahan (NASA GSFC)

arbitrary because it does not explain *why* the force of gravity is inversely proportional to the second power of the distance. It could have been proportional to some other power. In contrast, Einstein's theory gives you no freedom. His picture of gravity as curvature of spacetime, combined with the requirements of the principle of relativity, inevitably leads to Einstein's equations, which uniquely predict the inverse square law. In this sense, the general theory of relativity not only describes gravity, it *explains* gravity.

Summary

The key insight of Einstein's general theory of relativity (GR) is that gravity is a manifestation of spacetime curvature. Massive bodies curve the spacetime around them, causing nearby objects to move on curved trajectories. GR predicts that gravitational effects propagate at the speed of light. Thus, the force of gravity no longer acts instantaneously, as Newton had to assume. Nevertheless, for slow motions and weak gravitational fields, Einstein's theory reduces to Newton's inverse square law.

Since its inception, general relativity has been rigorously tested. The general relativistic calculation of the precession of Mercury's orbit is in perfect agreement with astronomical observations. The first new prediction of gen-

eral relativity to be tested observationally was the deflection of light as it propagates through a curved spacetime region near a massive body. Other important predictions include gravitational lensing, gravitational waves, gravitational time dilation and black holes. All of these effects have been observed.

Questions

1. (a) Imagine you are Sir Isaac Newton and you are wondering what would happen to the Earth's trajectory if the Sun instantaneously disappeared. Please describe.
(b) Now imagine that you are Albert Einstein and you are wondering what would happen to the Earth's trajectory if the Sun instantaneously disappeared. Please describe.
2. How is Newton's theory of gravity fundamentally incompatible with Einstein's special relativity?
3. Can you give an example of a 2-dimensional space that is homogeneous but not isotropic?
4. Consider a square tub of ice-cream. If we scoop out a perfect ball of ice-cream, is the ball an example of a three-dimensional curved space? (Hint: What do the angles of triangles add up to inside the ball of ice-cream?) Is the surface of the ice-cream ball an example of a two-dimensional curved space?
5. As we discussed in this chapter, Euclid's fifth axiom is violated in spherical and hyperbolic spaces. Now, consider Euclid's second axiom: "Any straight line segment can be extended without limit in either direction." Does it apply in spherical or hyperbolic space?
6. In a flat space, the apparent size of objects decreases with the distance. How would the apparent size of objects vary with the distance in a spherical space?
7. Suppose you are in a closed spherical universe, with stars uniformly spread over space. Is there an Olbers' paradox in such a universe? What observations would you do to test the hypothesis that you are in a spherical universe?
8. Consider a GPS clock in orbit around the Earth. Both special relativistic and general relativistic effects will alter the rate at which the clock ticks. These effects go in opposite directions. Explain.
9. Under what conditions does Einstein's GR reduce to Newtonian gravity? Is it a good or bad thing for GR to reduce to Newtonian gravity under the conditions you just listed? Why?
10. What is a gravitational lens?

11. Consider a falling elevator, as illustrated in Fig. 4.2. If the gravitational field is perfectly uniform, all objects in the elevator will fall with exactly the same acceleration.
 - (a) If the observer cannot see what is going on outside, can he perform any experiment that would distinguish being inside a falling elevator from being in an inertial frame of reference?
 - (b) Suppose now that the elevator is falling in the gravitational field of the Earth, which is not perfectly uniform. What experiment would you suggest to detect the presence of this gravitational field?
12. What is a gravitational wave? At what speed do gravitational waves propagate? Have such waves been detected?
13. Why did Einstein call his theories “special” and “general” relativity?

5

An Expanding Universe

5.1 Einstein's Static Universe

Shortly after completing the general theory of relativity, Einstein went on to apply his new theory to the universe as a whole. The structure of the universe beyond our Milky Way galaxy was then completely unknown, so Einstein had to make some assumptions. Following Newton, he assumed that on average matter is uniformly distributed in the cosmos. There are of course local variations, with the density of stars higher in some places and lower in others. However on very large scales the universe is well approximated as being perfectly homogeneous.

Einstein also assumed that the universe is isotropic on average. This means that it looks more or less the same in all directions. A homogeneous and isotropic (on average) distribution is illustrated in Fig. 5.1(a), with galaxies represented by dots. An example of a homogeneous distribution that is not isotropic is shown in Fig. 5.1(b), where the galaxies (dots) form a regular lattice. This distribution looks the same from every galaxy, but it looks different in horizontal, vertical, and diagonal directions. The actual distribution of galaxies, as revealed by modern astronomical observations, is more complicated than that in Fig. 5.1(a). Individual galaxies form clusters, which are in turn grouped into huge superclusters, typically 150 million light years across. But that is where the hierarchy of cosmic structure appears to end. If the distribution of galaxies is smoothed over distances of 300 million light years or so, it does appear to be homogeneous and isotropic.

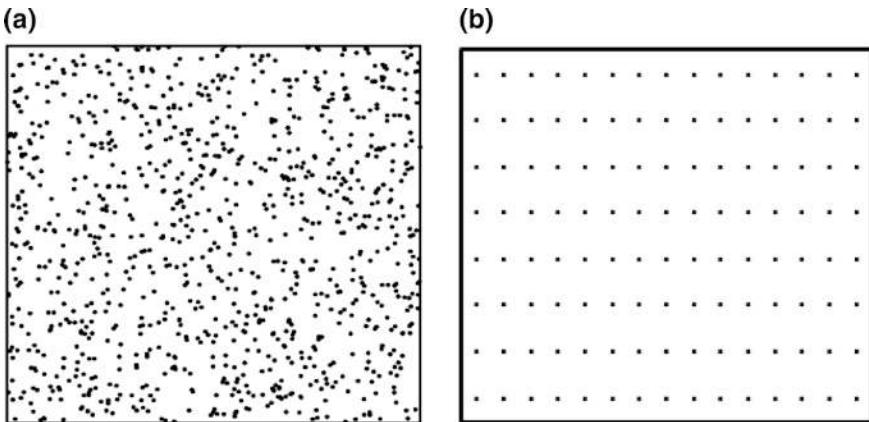


Fig. 5.1 **a** A homogeneous and isotropic (on average) distribution of galaxies.
b This distribution is homogeneous, but not isotropic

The universe cannot be homogeneous and isotropic unless space itself has these properties. The curvature of space should be (on average) the same at all places and in all directions. As we discussed in Chap. 4, there are only three types of such spaces: a flat Euclidean space, a closed spherical space, and an open hyperbolic space. A homogeneous and isotropic universe should therefore have one of these three geometries.

Finally, Einstein assumed that the average characteristics of the universe, such as the average density of stars, do not change with time. His overall picture was thus that the universe looks more or less the same at all places, in all directions, and at all times. Einstein did not have much observational data to back up his assumptions, but philosophically he found this picture of a homogeneous, isotropic, static universe very attractive.

It turned out, however, that the equations of GR have no solutions with these properties. The problem is that masses distributed in the universe are pulled together by gravity and refuse to stay at rest. The theory seemed to suggest that the universe could not be static. But the preconception of an eternal, immutable universe was too deeply rooted. Reluctantly, Einstein concluded that the equations of GR had to be modified, by adding an extra term, to allow for the existence of a static world.

The effect of the new term was to endow the vacuum—that is, empty space—with energy and pressure. This may sound crazy, but we know that Einstein was not afraid of making counter-intuitive assumptions and following them to their logical conclusion. According to the modified equations, the energy density of the vacuum ρ_v is constant everywhere; Einstein called

it the *cosmological constant*. The vacuum pressure P_v is related to the vacuum energy density ρ_v simply as

$$P_v = -\rho_v \quad (5.1)$$

Therefore, if ρ_v is positive, the pressure is negative. (We give a derivation of Eq. (5.1) from the work-energy relation at the end of this section.)

What does it mean for the pressure to be negative? The usual, positive pressure is an outward-pushing force, like the pressure of air in a balloon. Negative pressure is what we ordinarily call *tension*. It pulls inward, like the tension in a stretched piece of rubber. So, if the vacuum has tension, why does it not suck itself in and shrink? The reason is that in order to produce a force you need a *difference* in pressure: a balloon will expand if you increase the pressure inside it, but there will be no effect if the exterior pressure is increased by the same amount. The vacuum pressure is the same everywhere, and thus we should not expect any shrinkage (or expansion). The energy of the vacuum is equally elusive. There is no way to extract this energy; unfortunately we cannot solve the world's energy crisis by harnessing energy from empty space. The energy and pressure of the vacuum are thus completely unobservable—except for their gravitational effects.

The force of gravity in GR depends both on the energy (or mass) density, ρ , and the pressure, P . It is proportional to

$$\rho + 3P. \quad (5.2)$$

For ordinary matter, pressure is negligible, so we are used to thinking about gravity as being dependent only on mass. However, for the vacuum, the pressure has the same magnitude as the energy density (see Eq. (5.1)), and we find that the gravitational force of the vacuum is proportional to

$$\rho_v + 3P_v = -2\rho_v. \quad (5.3)$$

The negative sign here (in contrast to the positive sign for regular matter) indicates that the gravity of the vacuum is *repulsive*.

Einstein realized that by adding a cosmological constant to his equations, he could balance the gravitational attraction of matter with the gravitational repulsion of the vacuum. All he needed was a matter density with $\rho_m = 2\rho_v$, to perfectly balance the gravitational effect of the vacuum given in Eq. (5.3). He thereby obtained a solution that describes a static universe. This solution has a closed, spherical geometry, with the radius determined by the matter density. For the density given by recent measurements, the corresponding circumference is about 100 billion light years.

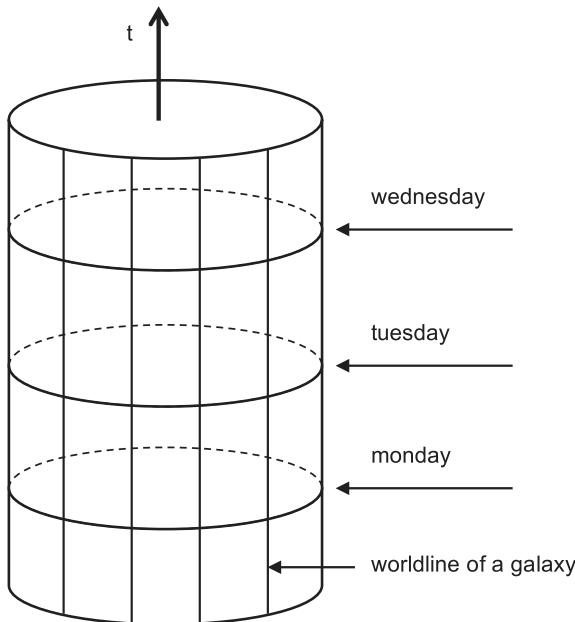


Fig. 5.2 Spacetime diagram of Einstein's static universe. Horizontal *circles* represent momentary snapshots of the universe. Two out of the three spatial dimensions are not shown

The spacetime of Einstein's static universe is illustrated in Fig. 5.2, with two out of the three spatial dimensions suppressed. It looks like the surface of a cylinder embedded in a 3-dimensional space, but only points on the surface belong to the spacetime. Time runs in the vertical direction, and horizontal slices give “snapshots” of the universe at different moments of time. In the figure these slices are circles, but in the four-dimensional spacetime the slices would be three-dimensional spherical spaces. The vertical straight lines are the worldlines of galaxies. In this universe, nothing changes with time, so all snapshots are identical and the positions of the galaxies do not change.

5.2 Problems with a Static Universe

Despite its philosophical appeal, it turns out that Einstein's static cosmological model is not acceptable. To see why, think about what would happen to the matter density, ρ_m , and the vacuum energy density, ρ_v , if the radius of the universe were slightly decreased. It doesn't take an Einstein to realize that

ρ_m would increase and ρ_v , by its very definition, would remain constant. This causes the balance scales to tip in the gravitational tug-of-war between the attraction of matter and the repulsion of the vacuum. Attraction prevails and the universe begins to contract. As it contracts, the matter density is further increased, so the contraction accelerates.

Similarly, we could ask: What would happen if the radius of the universe were increased slightly? In this case, ρ_m decreases, and the gravitational repulsion of the vacuum will win, causing the universe to expand ad infinitum. Small fluctuations in the radius of the universe cannot be avoided, and thus Einstein's universe cannot remain static for an infinite time.

Another problem with the idea of an eternal universe is that it is in conflict with one of the most universal laws of Nature—the second law of thermodynamics. This law states that an isolated physical system evolves from more ordered to more disordered states.¹ A gust of wind will lift papers from your desk and scatter them randomly over the floor, but you never see the wind picking up papers from the floor and organizing them neatly on the desk. A spontaneous ordering of this kind is not impossible in principle, but it is so unlikely that it is never seen to occur. A book sliding along the floor comes to a halt due to friction, and the energy of its directed, ordered motion turns into heat, that is, into the energy of the disordered motion of molecules. The inverse process would be for the book to cool down and start moving along the floor. This is forbidden by the second law of thermodynamics.

A mathematical measure of disorder is called *entropy*: the more entropy an object has, the more disordered it is. The second law says that the entropy of an isolated system can only increase. The evolution from ordered to more disordered states leads eventually to the state of maximum entropy, known as *thermal equilibrium*. In this state, all ordered motion ceases, all energy is converted into heat, and a uniform temperature is established throughout the system.

The universe can be regarded as an isolated system (since there is nothing outside of it). Therefore, if it existed forever, thermal equilibrium would have already been reached. The stars would have completely burnt out, cooling to the same temperature as interstellar space, and no life would be possible.² But this is not what is observed, so the universe could not have existed forever.

¹In case you are wondering, the first law of thermodynamics is just a statement of energy conservation, generalized to include thermal processes. It states that the total energy of an isolated system, including its heat energy, is conserved.

²This bleak prediction was publicized by the German physicist Hermann von Helmholtz. He called it the “heat death” of the universe.

There is a caveat though. The Austrian physicist Ludwig Boltzmann realized that even in thermal equilibrium, spontaneous reductions of disorder occasionally happen by chance. They are called *thermal fluctuations*. So in order to reconcile the second law of thermodynamics with a universe that has existed forever, and an observable universe that is not in thermal equilibrium, we would have to conclude that we are living in a huge thermal fluctuation.

You may be concerned that such a huge fluctuation is extremely improbable. True. But if life can only exist in ordered parts of the universe, one can argue that this explains why we are observing such an incredibly rare event. Yet if we take this approach, we are still at a loss to explain why we don't find ourselves in a much smaller, and much more probable, fluctuation. It would suffice to turn chaos into order on the scale of the Solar System as opposed to the vastly larger scale of the observable universe.

Derivation of Eq. (5.1)

Consider a chamber of volume V filled with a vacuum of energy density ρ_v and pressure P_v . The volume of the chamber can be varied by moving a piston, as shown in Fig. 5.3. The total energy of the vacuum is

$$E = \rho_v V, \quad (5.4)$$

and the force it exerts on the piston is

$$F = AP_v, \quad (5.5)$$

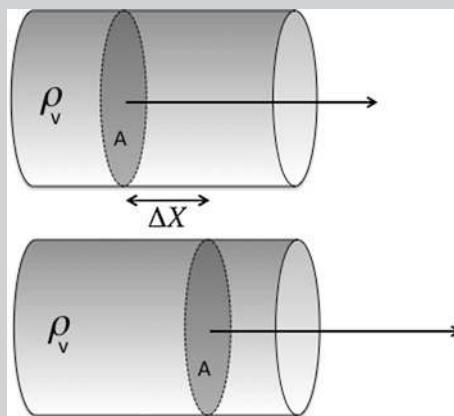


Fig. 5.3 Changing the volume of a chamber filled with constant energy density

where A is the surface area of the piston. (Recall that pressure is the force per unit area, $P = F/A$.) Suppose the piston is moved outwards by amount Δx , so that the volume is increased by

$$\Delta V = A\Delta x. \quad (5.6)$$

You may remember from elementary physics that the resulting change in the energy is

$$\Delta E = -\Delta W, \quad (5.7)$$

where ΔW is the work, which is defined as

$$\Delta W = F\Delta x. \quad (5.8)$$

(The work is positive if the force is in the direction of motion of the piston and negative otherwise.)

Thus, using Eqs. (5.7), (5.8), (5.5), and (5.6), we can show that the change in the energy of the vacuum is

$$\Delta E = -F\Delta x = -P_v A\Delta x = -P_v \Delta V. \quad (5.9)$$

Using $\Delta E = -P_v \Delta V$ (from Eq. (5.9)), and using $\Delta E = \rho_v \Delta V$ (from Eq. (5.4)), we find that the pressure of the vacuum is related to its energy density as $P_v = -\rho_v$.

5.3 Friedmann's Expanding Universe

The next breakthrough development in cosmology occurred in a rather unlikely place—the Soviet Petrograd, devastated by war and the Russian revolution. It took several years for Einstein's papers on GR to reach Russia. Once they got there, the young mathematician Alexander Friedmann voraciously studied the theory, focusing on what he thought was its central problem—the structure of the universe as a whole. He adopted Einstein's assumptions that the universe was homogeneous and isotropic and that it had a closed spherical geometry. Then he took a radical step: he did not require that the universe is static (Fig. 5.4).

With the requirement of a static universe lifted, Friedmann found that Einstein's equations did have a solution. The solution describes a spherical universe with a time-varying radius and mass density. It starts with zero radius, expands, comes to a halt, and then contracts back to size zero. If you were an observer living in some galaxy in such a universe, then during the expansion phase, you would see all the other galaxies moving away from your galaxy, whilst during contraction all galaxies would approach you. It might seem that you are located at some special cosmic center, but observers in all the other galaxies would see the same thing.



Fig. 5.4 The Russian mathematician Alexander Friedmann (1888–1925) was the first to find time-dependent solutions of Einstein's equations, describing an evolving cosmos. During World War I, while Einstein was completing his general theory of relativity, Friedmann served as a bomber pilot in the Russian air force. He was awarded a George Cross for bravery. Apart from his work in cosmology, Friedmann did groundbreaking research in hydrodynamics and meteorology. He died of typhoid fever at the age of 37

To understand how this is possible consider the surface of a balloon, which is a good 2D analogy to a 3D closed spherical geometry. Imagine that small dots, representing galaxies, are painted on the balloon (see Fig. 5.5). As the balloon is inflated, all the dots move away from each other as their relative distances increase. Conversely, as the balloon is deflated, all the dots get closer to each other. It doesn't matter which dot we focus on, the view is the same.

One limitation of this 2D analogy is that when a balloon expands, it expands into the volume of air surrounding it. So, what does Friedmann's universe expand into? Nothing. In the analogy, the surface of the balloon is all the 2D space there is. The amount of space (the area) grows as the balloon expands—but there is nothing outside or inside the surface. Similarly, the total volume of a 3D Friedmann universe grows during the expansion and decreases during the contraction.

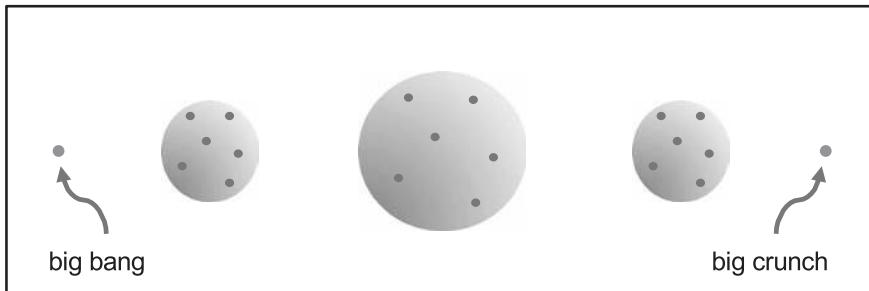


Fig. 5.5 The universe begins as a point and expands until gravity finally halts the expansion, and the universe collapses back to a point

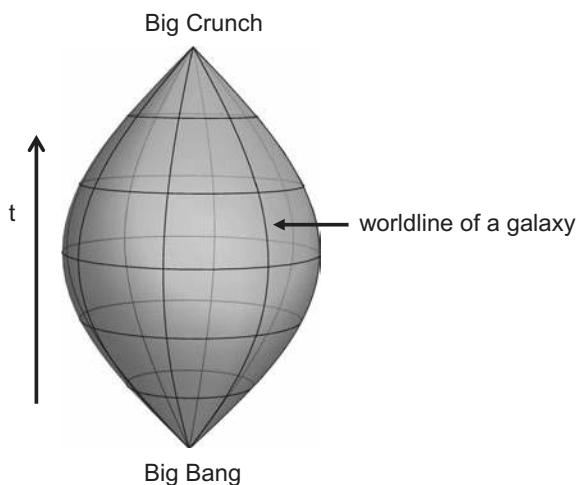


Fig. 5.6 Spacetime diagram of a closed universe. Horizontal circles are momentary snapshots of the universe, and the "meridian" lines are worldlines of galaxies

The history of an evolving closed universe is encapsulated in the spacetime diagram in Fig. 5.6. Here, time runs from the bottom up and horizontal circles represent instantaneous snapshots of the universe (with two spatial dimensions suppressed). The initial and final moments were later named, somewhat disrespectfully, the *big bang* and the *big crunch*. At these moments, all matter is compressed into an infinitesimal volume (a single point), so the density is infinite. This makes Einstein's equations mathematically ill defined, so the spacetime cannot be extended beyond these points. Such points are called spacetime *singularities*.

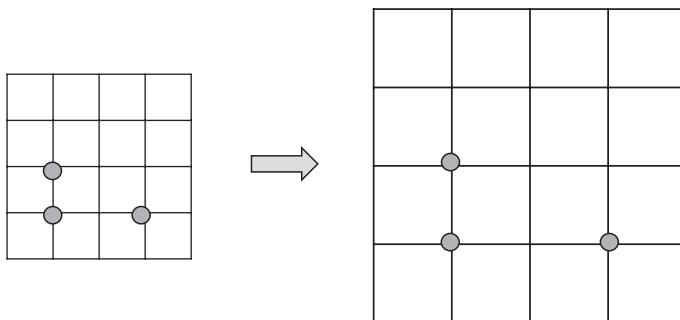


Fig. 5.7 2D stretched rubber sheet with galaxies represented by *circles*

According to Friedmann's solution, shortly after the big bang, the expansion of the universe is very fast. Then it is slowed down by the gravitational attraction between galaxies, and eventually comes to a halt, followed by contraction. This is similar to the motion of a projectile launched vertically upwards. The projectile is slowed down by gravity until it reaches some maximum height and then falls back to the ground. The greater the initial velocity, the higher it will go. Similarly, a Friedmann universe will expand to a larger radius if the initial expansion rate is increased.

Friedmann presented his solution in a paper that was published in 1922 in a German physics journal. Two years later, he published a follow-up paper describing an infinite (open) homogeneous and isotropic universe with hyperbolic geometry. Once again, he found that such a universe expands from a singularity of infinite matter density. The expansion slows down initially but it never stops completely, with galaxies approaching constant recession speeds at late times. This is analogous to a projectile launched at a speed exceeding the escape velocity (see Chap. 2). The gravitational pull of the Earth is not strong enough to turn it around, and the projectile permanently leaves the Earth.

The borderline case between open and closed solutions is a “flat” universe, having Euclidean geometry. Such a universe expands forever, but at ever decreasing speed, like a projectile launched at exactly the escape velocity.³ A 2D analogue for an expanding flat universe is a flat rubber sheet that is being uniformly stretched in both directions. The distances between all “galaxies” are then stretched by the same factor (see Fig. 5.7). The sheet can be arbitrarily large, and we can imagine it to be infinite. When we say that a flat universe has expanded by a certain factor, what we mean is that distances between all galaxies have increased by that factor.

³Friedmann did not consider this borderline case. It was later studied by Einstein and Willem de Sitter.

Friedmann did not give preference to closed or open universe models. He wrote: "The available data is completely insufficient for any numerical estimates to find out what kind of universe is ours." Sadly, Friedmann died in 1925, before his papers had attracted much attention. The Belgian priest Georges Lemaître rediscovered the expanding universe models in 1927, but his work also passed unnoticed. All this changed in 1929, when Edwin Hubble made what was arguably the most unexpected discovery in the history of science: he observed that the universe is indeed expanding! Friedmann and Lemaître were immortalized (Fig. 5.8).

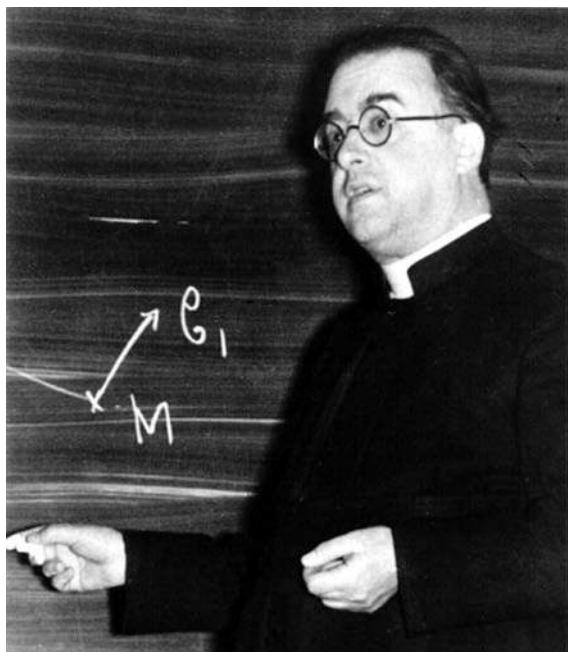


Fig. 5.8 Georges Lemaître (1894–1966) interrupted his undergraduate studies to serve in the Belgian army during World War I. After the war he went back to university and earned a Ph.D. in mathematics in 1920. He then went on to study for the priesthood, becoming ordained in 1923. By this time Lemaître developed an interest in astronomy, which he pursued at Cambridge, Harvard, and then at MIT, where he earned his second Ph.D. In his Ph.D. thesis Lemaître rediscovered Friedmann's solutions of Einstein's equations describing an expanding universe. He also showed that recession speeds of galaxies in such a universe should obey what is now known as the Hubble law—two years before Hubble's discovery. Lemaître explained his ideas to Einstein at a conference in Brussels in 1927—to which Einstein replied: "Your calculations are correct, but your grasp of physics is abominable." A few years later Einstein changed his mind. Being both a Catholic priest and a renowned scientist, Lemaître saw no conflict between science and religion. He believed that religion should keep to the spiritual world, leaving the material world for science.

As for Einstein, he reportedly quipped that adding the cosmological constant to his equations “was the greatest blunder of my life”. But even though the cosmological constant fell out of favor after the expansion of the universe was observed, it has since returned to the forefront of physics research, and we shall have much more to say about it in the coming chapters.

Summary

As soon as Einstein had completed the general theory of relativity, he applied it to the universe as a whole. Like Newton, Einstein believed that the universe was static and eternal, but he soon discovered that his theory did not admit such solutions. He then added an extra term to his equations, the so-called cosmological term, which endowed the vacuum with a non-zero (positive) energy density. According to general relativity, the vacuum then produces a repulsive gravitational force, which can balance the attractive gravity of matter. The modified equations had a static solution, describing a closed, spherical universe, but this model was seriously flawed. It was unstable to small perturbations and contradicted one of the most fundamental laws of Nature—the second law of thermodynamics.

In the meantime, the Russian mathematician Alexander Friedmann found dynamical solutions of Einstein’s equations describing evolving universes that expand from a singular state of infinite density. His closed geometry solution describes a finite universe that starts out expanding rapidly, slows down, and eventually turns around and starts to collapse. The open geometry solution describes an infinite universe that starts out expanding rapidly, and although the expansion slows down, it never stops completely. Flat expanding universes are the marginal case, between open and closed. They are infinite and galaxies approach a recession speed of zero.

Questions

- When Einstein first applied GR to the universe as a whole he assumed that the universe is homogeneous and isotropic. This is known as the “cosmological principle”. Is this principle consistent with a universe which has a center or an edge?
- Is the distribution of galaxies in Fig. 5.9 homogeneous? Is it isotropic about the central galaxy? Is it isotropic about any other galaxy?
- True or false: If the universe is isotropic about every galaxy, it must also be homogeneous.
- Why did Einstein add a cosmological constant to his equations of GR?

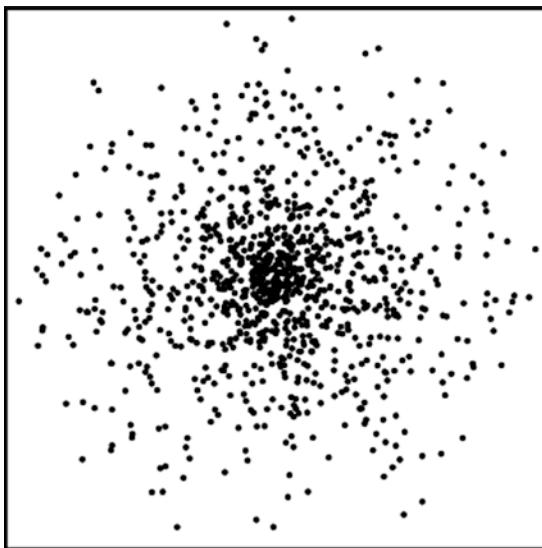


Fig. 5.9 A distribution of galaxies

5. Einstein included a positive cosmological constant ($\rho_v > 0$) in his equations. What would have happened if he had added a negative cosmological constant?
6. Einstein's cosmological constant endows the vacuum with negative pressure. Negative pressure acts like tension in a piece of rubber. So why doesn't the universe suck itself in? What effect does negative pressure have on the expansion rate of the universe?
7. What do we mean when we say Einstein's static model of the universe is unstable?
8. Is the following a correct statement of the second law of thermodynamics: "Any physical system evolves from more ordered to more disordered states"? If not, why not?
9. Why is Einstein's model of the universe in conflict with the second law of thermodynamics?
10. In a spacetime diagram for a static Einstein universe, like the one in Fig. 5.2, sketch the worldline of a flash of light emitted from some galaxy, which runs around the universe and returns to the same galaxy.
11. For a Friedmann closed universe, what is the density and radius of the universe at $t = 0$, the time of the big bang? Are the equations of general relativity valid at $t = 0$?
12. What is a spacetime singularity?

13. We used two-dimensional balloon and rubber sheet analogies to visualize closed and flat expanding universe models. Can a similar visualization be set up for an open, hyperbolic universe?
14. Is it possible to distinguish inertial motion from rest in Einstein's universe? In other words, is there any special class of inertial observers in such a universe, which can be characterized as being at rest?
15. Consider two twins who live in Einstein's static closed universe. One of the twins sets out in a rocket and heads away from his sibling at near light speed. Eventually the travelling twin returns to where he started due to the curvature of space. As he passes his twin and looks at him, which of them is older? (Note they both have maintained only inertial motion).

6

Observational Cosmology

Giordano Bruno was burned at the stake for his heretical ideas in 1600. He believed that the stars are like our Sun and appear to be dimmer only because of their great distance from us. This was an inspired guess, but how can we verify that it is actually true? How far away are the stars? And what are they made of?

These questions bedeviled Isaac Newton. He thought there was a distinction between the “lucid matter” of the stars and the “opaque matter” of the Earth and the planets. In a letter to Richard Bentley, where he discussed the creation of the Solar System amongst other things, Newton wrote: “But how the matter should divide itself into two sorts, and that part of it which is to compose a shining body should fall down into one mass and make a sun and the rest which is fit to compose an opaque body should coalesce, not into one great body, like the shining matter, but into many little ones; or if the sun at first were an opaque body like the planets or the planets lucid bodies like the sun, how he alone should be changed into a shining body whilst all they continue opaque, or all they be changed into opaque ones whilst he remains unchanged, I do not think explicable by mere natural causes, but am forced to ascribe it to the counsel and contrivance of a voluntary Agent.”

While Newton resorted to divine intervention to explain the separation of the Solar System into the lucid Sun, and opaque planets, spectroscopic experiments in the mid 1800s revealed that the Sun and the stars are actually made

of the same chemical elements as the Earth and planets.¹ In this chapter we will study how spectroscopy allows us to identify chemical elements; even those in distant stars. We will also learn how the Doppler effect is used to measure velocities of cosmic bodies, and how astronomical distances are determined.

6.1 Fingerprints of the Elements

Light coming to us from the stars brings a treasure-trove of information. We learned in Chap. 3 that light consists of electromagnetic waves that can have a wide range of wavelengths (or frequencies). For visible light, different wavelengths correspond to different colors. When a beam of white light is shone through a prism, it emerges on the other side having a continuum of colors, like a rainbow (see Fig. 6.1a). This *continuum spectrum* shows that white light is composed of many colors, ranging from red to violet.

Light emitted from a hot gas that is incident on a prism displays an *emission* line spectrum—a pattern of bright lines with particular wavelengths can be seen on a black background (Fig. 6.1b). Another interesting phenomenon occurs when a beam of white light is passed through cool gas before it gets to the prism. The gas absorbs waves having some specific wavelengths, and a pattern of black lines, called *absorption lines*, appears in the spectrum (Fig. 6.1c). The pattern of both emission and absorption lines depends on the composition of the gas. Atoms of a given chemical element can emit and absorb light only at a particular set of wavelengths,² so the emission or absorption line spectrum provides a unique “fingerprint” for each element.

Light that emanates from the hot inner parts of a star has a continuum spectrum³ that develops absorption lines as it passes through the cooler stellar atmosphere. Astronomers measure the spectra of stars, and by comparing with the absorption lines of gases measured in the laboratory, they can identify whether elements such as hydrogen, helium, carbon, etc., are present in the star. In fact, helium was discovered on the Sun in 1868, well before it was found on the Earth in 1895. Stellar spectroscopy has indisputably determined that stars are indeed made of the same “stuff” as the Earth.

¹The development of nuclear physics has led to a detailed understanding of how “opaque matter” can become “lucid matter” under the right conditions. Even more astonishing is that most of the elements from which we are made were actually produced in the stars themselves (as we will discuss later).

²This fact was already known in the mid-1800s, but was explained only much later by quantum mechanics.

³Although light emitted by atoms has a discrete line spectrum, in stellar interiors atoms are broken up into electrons and nuclei which scatter off one another, producing a continuous spectrum.

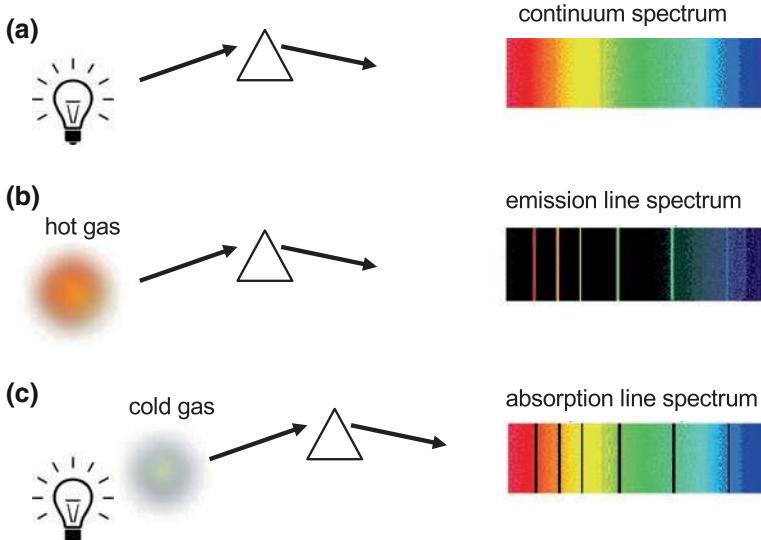


Fig. 6.1 **a** When white light passes through a prism, it spreads into a continuum of colors. **b** A hot gas emits specific wavelengths that show up as bright lines on a black background. **c** A cold gas absorbs specific wavelengths that are then absent from the continuum spectrum. Notice that the emission lines have the same wavelengths as the absorption lines (as long as the hot and cold gas are of the same type)

6.2 Measuring Velocities

For nearby stars, like Barnard's star, we can directly calculate how fast the star is moving in a direction orthogonal to the line of sight. We do so by measuring the star's displacement on photographic plates taken some time apart. However, more distant stars are so far away that it is impossible to detect their motion and measure their velocities using this method. So how do we measure the velocity of astronomical objects? (Fig. 6.2).

The observed wavelength (color) of light depends on the relative motion of the source and the observer. If a source of light is moving towards us, the observed wavelength gets shorter—that is, it shifts towards the blue end of the spectrum. Conversely, if a source is moving away from us, the observed wavelength will get longer, shifting towards the red end of the spectrum. We say the source is blue- or redshifted. This phenomenon, known as the *Doppler effect*, occurs for all kinds of waves, including sound waves and ripples on the surface of water. You have probably experienced it when a siren has passed by: an approaching siren has a higher pitch (shorter wavelength) than a receding one. As illustrated in Fig. 6.3, the wave crests pile up in

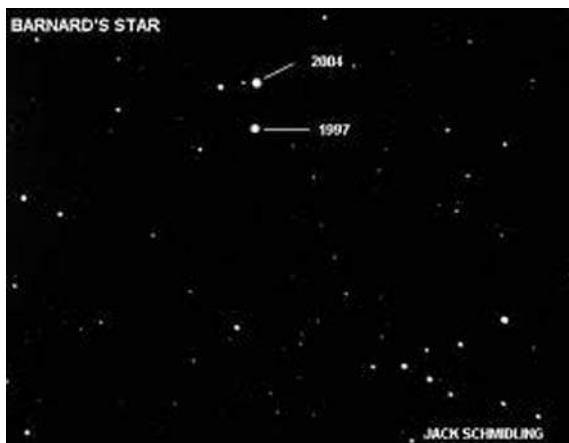


Fig. 6.2 Barnard's star, shown here at two different times, is about 6 light years away. Credit © Schmidling Productions "Barnard's star" (Encyclopædia Britannica Online. Web. 24 Dec. 2016)

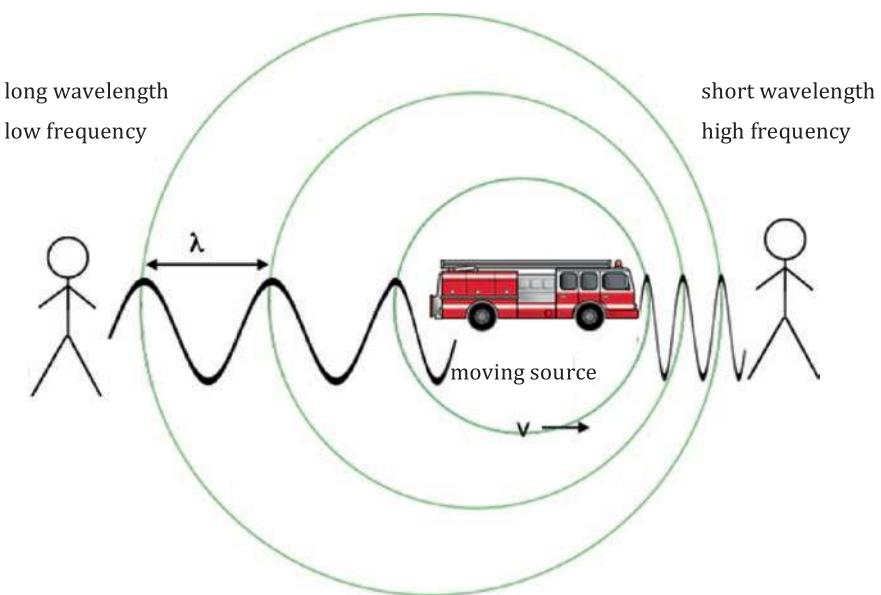


Fig. 6.3 Doppler effect for sound. Approaching sirens have a higher pitch than receding ones Credit NASA's Imagine the Universe

front of a moving source and spread out in its wake. Remembering that the wavelength is the distance between the crests, it is easy to see that the wavelength gets shorter in front and longer in the wake of the source.

Quantitatively, the Doppler effect for light can be expressed by a simple formula:

$$\frac{\Delta\lambda}{\lambda} = \frac{v}{c} \quad (6.1)$$

where λ is the wavelength, v is the relative velocity of the source and observer, and c is the speed of light. As before, the symbol Δ stands for “change”, so $\Delta\lambda$ is the change in the wavelength λ . The velocity v is assumed to be small compared to c (nonrelativistic motion), and it is taken to be negative if the source is approaching and positive if it is receding. (Note: λ is the emitted wavelength, and $\Delta\lambda = \lambda_o - \lambda$, where λ_o is the wavelength measured by the observer.) The *redshift* z is defined as

$$z \equiv \frac{\Delta\lambda}{\lambda}. \quad (6.2)$$

Thus, using Eq. (6.1), we note that for nonrelativistic motion, $z = v/c$.

For a moving star, the entire spectrum gets blue- or redshifted, including the black absorption lines. Astronomers identify line patterns of different elements and measure how much these patterns are shifted relative to a sample at rest in the laboratory. Equation (6.1) can then be used to determine the velocity of the star.⁴ It is hard to overstate the crucial role spectroscopy and the Doppler effect play in our endeavor to understand the universe.

6.3 Measuring Distances

The determination of distances to astronomical objects is notoriously difficult and has dominated much of twentieth century astronomy. Today astronomers use a variety of techniques to measure distances—each one is most useful within a given range. Distances to nearby stars can be found by measuring their *parallax*, which is the apparent movement of the star relative to the background sky as the Earth rotates around the Sun (see Fig. 6.4). Demonstrating the effect of parallax is so simple that anyone can do it—you don’t even need a telescope! If you stretch out your arm, hold up

⁴Note that Doppler effect can be used only to measure velocities along the line of sight, that is, towards or away from us. Transverse velocities in the orthogonal directions cannot be measured in this way.

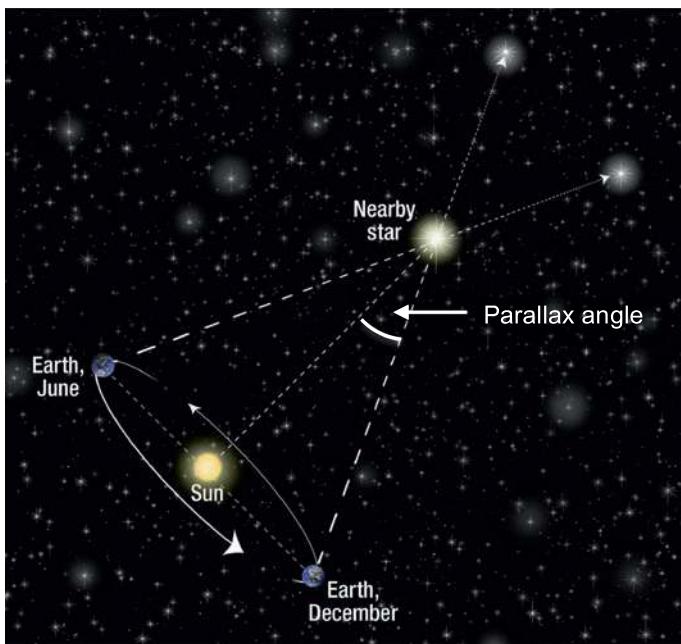


Fig. 6.4 The apparent shift in position of a nearby star relative to very distant background stars allows us to determine the nearby star's distance. The Earth's orbital diameter can be used as a baseline if we view the star at the beginning and end of a 6 month period. In reality (unlike the figure) the distance to the stars is much greater than the Earth's orbit, so the parallax angle is very small
Credit NASA, ESA, and A. Feild (STScI)

your thumb, and alternately close your right and left eyes, you will see that your thumb appears to alternate between two different positions relative to the back of the room. From some simple geometry, knowing the distance between your eyes (the “baseline”), and the angular shift of your thumb (twice the parallax angle), you can determine the distance to your thumb.

The parallax is used to define an astronomical unit of distance, called a *parsec* (pc). One parsec is the distance at which a star would have a parallax of $1''$.⁵ It is equal to about 3.3 light years. In this book we will usually express distances in light years, and not parsecs. Since parallactic angles are very small, it becomes extremely hard to measure them for objects that are more than about 100 light years away.

⁵An arc second is a measure of angle. There are 360° in a full circle, $60'$ in a degree, and $60''$ in an arc minute. An arc second is a tiny angular measure (it is about the angle subtended by a dime placed 4 km away).

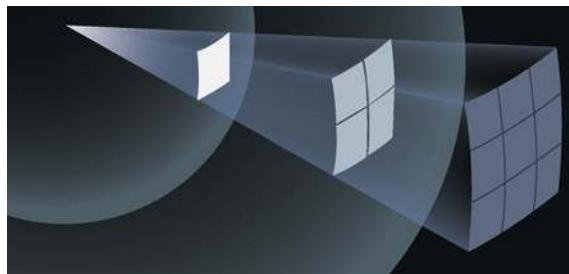


Fig. 6.5 The energy emitted by the source is spread over a *spherical* surface, whose area grows like the *square* of the distance from the source

While 100 light years seems like a large distance, our nearby neighbor, the Andromeda galaxy, is 2.5 million light years away. So parallax measurements can be seen as a first rung in what we call the cosmic distance ladder. Astronomers use a variety of so called *standard candles* to extend the reach of our distance measurements. Although none of them is perfect, they all work on the following premise: if we know how intrinsically luminous a light source is, and we measure how bright it appears, we can figure out how far away it is. The key relation is that the brightness of a light source decreases with the square of its distance,

$$b = \frac{L}{4\pi d^2} \quad (6.3)$$

The luminosity L is the energy of light emitted by the source per second. As the light travels a distance d from the source, this energy gets spread over a sphere of area $4\pi d^2$, and the apparent brightness b decreases accordingly (see Fig. 6.5).

Pulsating stars, called Cepheids, are particularly useful standard candles. Their brightness varies periodically, with periods ranging from days to months. A remarkable property of Cepheids, discovered in 1912 by Henrietta Leavitt of Harvard College Observatory, is that they display a tight relationship between their period of variation (which is easy to measure) and their luminosity, as shown in Fig. 6.7. Thus by measuring the period, we can deduce the luminosity L . We can also measure the apparent brightness b , and once we know L and b , we can use Eq. (6.3) to determine the distance to the star. Cepheids can be used to measure distances up to about 10 million light years (Fig. 6.6).

Today astronomers use extremely powerful stellar explosions, called *supernovae*, as standard candles. Although there are many kinds of supernovae, with differing properties, *Type Ia supernovae* have very uniform luminosities and



Fig. 6.6 Henrietta Swan Leavitt (1868–1921) received an excellent education from Radcliff College, but being a woman she was unable to work as an official academic. Instead she found work as a “human computer” (with many other women) at Harvard College Observatory, where she earned the equivalent pay of a servant. She was a quiet, hard working woman, whose seminal discovery of the period-luminosity relationship for Cepheid stars made it possible for astronomers to measure the Universe. Despite the importance of her discovery, Leavitt received almost no credit in her lifetime. A member of the Swedish Academy of Sciences tried to nominate her for the Nobel Prize in 1924, only to discover that she had died of cancer three years earlier, at the age of 53

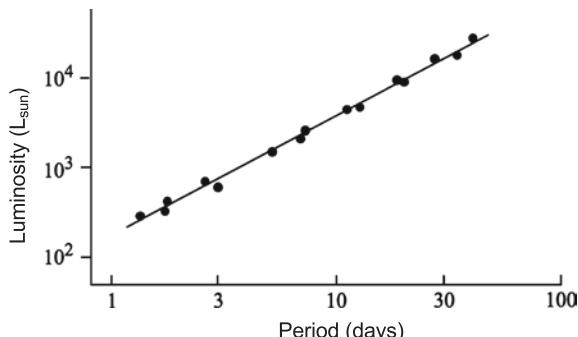


Fig. 6.7 A sketch of the period-luminosity relation for Cepheid variable stars
Credit Mark Whittle

are thus excellent standard candles. The physics of Type 1a supernovae is not yet fully understood, but the most plausible cause is a thermonuclear explosion of a white dwarf star.⁶ There appear to be two mechanisms to trigger the explosion. Firstly, if a white dwarf has a companion star, from which it can

⁶When an ordinary star (with a mass similar to the Sun’s) depletes its nuclear fuel, it becomes a very dense compact white dwarf star. The pull of gravity in a white dwarf is balanced by the pressure of the material within the star.

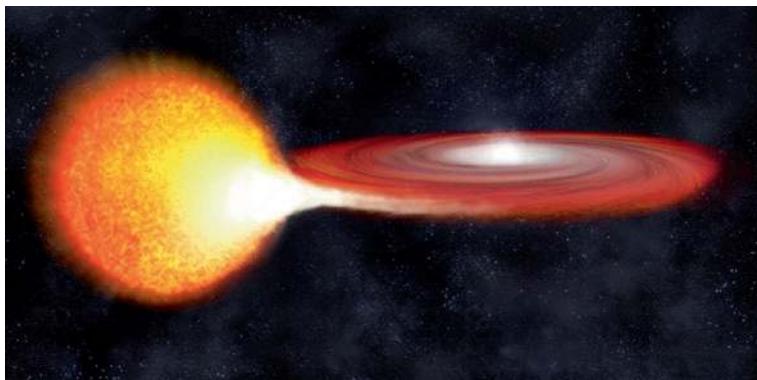


Fig. 6.8 Artist's impression of a white dwarf star accreting matter from a binary companion. When the star reaches a certain mass threshold it explodes, becoming a supernova Credit NASA/CXC/M.Weiss

accrete material, it may gain so much mass that gravity overwhelms the pressure forces, and the white dwarf starts to collapse. This ignites a runaway thermonuclear reaction, and the white dwarf star is completely blown away. An alternative scenario is a collision of two white dwarfs. When the two stars merge, their combined mass exceeds the stability threshold, and once again this leads to collapse. Whatever the mechanism, there is strong observational evidence that Type 1a supernovae have nearly the same peak luminosity. By measuring the apparent brightness of such supernovae and knowing the luminosity, the distance to the host galaxy can be determined. These powerful beacons have allowed astronomers to chart the universe out to billions of light years (Fig. 6.8).

6.4 The Birth of Extragalactic Astronomy

By the turn of the 20th century, astronomers had identified two types of objects outside our Solar System—point-like stars and faint, fuzzy extended objects called *nebulae*. The great question of the day was “*What is the nature of the nebulae?*” There were two rival theories. The first theory advocated that there was nothing but empty space beyond our Galaxy. Nebulae were considered to be objects within the Galaxy, probably sites of star formation. The opposing view held that nebulae were distant “island universes” in their own right, similar to our Galaxy. This contentious question resulted in “The Great Debate” between Harlow Shapley and Heber Curtis, held in 1920 at the Museum of Natural History in Washington. The debate ended inconclusively,

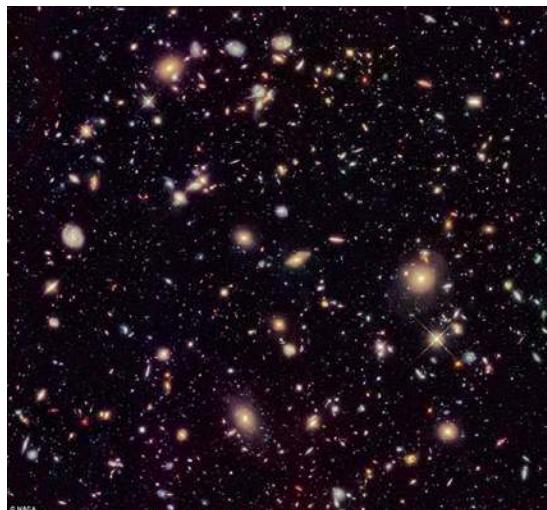


Fig. 6.9 Stars and nebulae *Credit NASA*

but the issue was definitively resolved in 1923, when Edwin Hubble established that the nebulae were other island universes, completely separate from our Galaxy (Fig. 6.9).

Hubble identified Cepheid variable stars in Andromeda and several other nebulae. Using Leavitt's period-luminosity relation, he was then able to determine the distances to these nebulae. Today we know that Andromeda is about 2.5 million light years away—roughly 50 times the radius of the Milky Way. Hubble's initial estimate of 1.5 million light years was significantly lower. However, it was still large enough to show that the nebulae must include billions of stars; and that they are indeed “island universes” similar to our own Galaxy. We now call them galaxies (Fig. 6.10).

Summary

Each chemical element displays a characteristic pattern of spectral lines. By analyzing the spectra of light coming from stars and galaxies we can determine their chemical composition. Furthermore, the spectral lines may be shifted relative to a laboratory sample here on Earth. From this shift we can determine velocities using the Doppler effect. Distances to nearby stars can be found using stellar parallax, while for more distant objects astronomers use a variety of “standard candles” like Cepheid stars and supernovae. In par-



Fig. 6.10 Pinwheel galaxy Credit ESA and NASA

ticular, Edwin Hubble used Cepheids to establish that the then mysterious spiral nebulae were not part of our Galaxy, but were separate distant galaxies.

Questions

1. What are emission and absorption line spectra?
2. Does red light have a longer or shorter wavelength than blue light? Does it have a higher or lower frequency than blue light?
3. If an object is approaching us, will its spectral lines be blue or red shifted? Explain.
4. An unshifted (laboratory) emission line spectrum of pure hydrogen (top), and an emission line spectrum from a moving object are shown in Fig. 6.11. Using the Doppler formula Eq. (6.1), calculate the velocity of the moving object. Is it moving toward or away from the observer?
5. The distances to nearby stars are found by measuring their parallax. If the parallax angle of star A is twice that of star B, which of the two stars is closer to us? By how much?
6. What is a “standard candle” and how do astronomers use them to measure distances?
7. Imagine that you have measured the distance to a galaxy using a standard candle. After you publish your results, it comes to light that your standard candle is twice as luminous as you had thought. How is the distance to the galaxy modified?



Fig. 6.11 Hydrogen emission line spectra. (Wavelengths are measured in nanometers.)

8. A 50 W light bulb is placed at a distance 10 m away, and a 100 W bulb is placed at a distance 20 m away. Which of the two bulbs appears brighter? By how much?
9. How can we use Cepheid variable stars to measure distances?

7

Hubble's Law and the Expanding Universe

In the early 1900s, Vesto M. Slipher of the Lowell observatory in Arizona analyzed the spectra of many spiral nebulae. He found that most of them mysteriously had spectral lines that were red-shifted, indicating that they were moving away from the Earth, some at speeds of up to 1000 km/s. Motion at high speed is not uncommon in the cosmos—the Sun, for example, orbits around the center of our Galaxy at 300 km/s. The puzzling thing about Slipher's result was that the nebulae conspired to move predominantly away from us, as if in a display of some cosmic rejection (Fig. 7.1).

Hubble set out to investigate Slipher's curious findings. He started by measuring distances to an extended sample of nebulae, now recognized as galaxies. Unfortunately, Cepheids were too faint to be observed in all but the nearest galaxies, so Hubble had to find a new standard candle. He noticed that the brightest stars, in those galaxies whose distances he could measure (using Cepheids), had about the same luminosity, so he used them as standard candles, extending the cosmic distance ladder. In the meantime, Hubble's assistant Milton Humason extended Slipher's redshift measurements to a larger set of galaxies. Hubble then plotted the redshifts obtained by Slipher and Humason versus his distance estimates. He published his findings in 1929—and our view of the universe has never been the same (Fig. 7.2).



Fig. 7.1 Vesto Slipher undertook the painstaking task of obtaining spectra for various spiral nebulae because he wanted to understand the origin of the Sun and planets. At the time, it was commonly thought that spiral nebulae might be other Solar systems in the process of forming

7.1 An Expanding Universe

Hubble uncovered a very simple relation between the speed at which a galaxy moves away from us and the distance to the galaxy¹: the speed grows proportionally to the distance. The further away the galaxy is, the greater is its speed. If you double the distance, the speed is also doubled. This is the celebrated Hubble law (Fig. 7.3).

Mathematically, the Hubble law can be stated as follows

$$v = H_0 d \quad (7.1)$$

¹More precisely, Hubble uncovered a linear relation between the redshift of a galaxy and its distance. The redshift is then converted to a recession velocity.



Fig. 7.2 Edwin Hubble (1889–1953) made some of the most important discoveries in modern astronomy. He showed that our Milky Way is only one of a multitude of galaxies scattered throughout the cosmos. His greatest achievement though was the discovery of the expansion of the universe. After graduating from Law school at Oxford University, and a short stint practicing Law and then teaching at a high school, Hubble obtained a Ph.D. in Astronomy at the University of Chicago in 1917. He was offered a job at Mt Wilson Observatory, which he took up only after first enlisting in the US Army to fight Germany. Hubble was a talented athlete, excelling in track and boxing amongst other sports. Had he not died suddenly from a stroke at the age of 63, he most probably would have been awarded a Nobel Prize, something that was impossible earlier in his career, as astronomers were then not eligible. Credit Hale Observatories, courtesy AIP Emilio Segrè Visual Archives

where v is the galaxy's velocity, and d is its distance. The constant of proportionality is called the Hubble parameter; its numerical value is²

$$H_0 = 2.2 \times 10^{-18} \text{ s}^{-1}. \quad (7.2)$$

²Because of uncertainties in distance measurements, it took scientists more than half a century to converge on this value: Hubble's original estimate was 4 times higher.

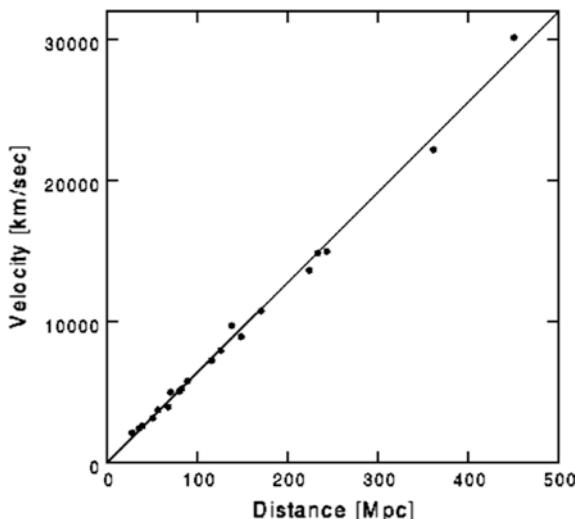


Fig. 7.3 Hubble's law with data from the High Redshift Supernova team (1996). Recession velocity is plotted versus the distance, measured in megaparsecs (Mpc) ($1 \text{ Mpc} \approx 3 \times 10^6$ light years). Credit Ned Wright (UCLA) using data from Riess, Press & Kirshner (1996, astro-ph/9604143)

At first sight it may appear that the Hubble law implies that we are located right at the center of some gigantic explosion. But the work of Friedmann and Lemaitre demonstrated that cosmic expansion need not have a center. In a homogeneous and isotropic expanding universe, all observers see the surrounding galaxies recede. Moreover, it is not difficult to understand that they must recede according to the Hubble law.

Once again, we can picture an expanding universe using the rubber sheet analogy (see Fig. 7.4). The sheet is uniformly stretched in both directions, and the dots on the surface of the sheet represent galaxies. Suppose, for the sake of argument, that the sheet has been stretched to twice its original size in one second. The dots that were initially 1 cm apart are now 2 cm apart, so they separated at the speed of 1 cm/s. At the same time, the dots that were 2 cm apart are now 4 cm apart and therefore separated at 2 cm/s—twice the speed of the first pair of dots. You can easily convince yourself that the speed at which any two dots separate is proportional to the distance between them. But this is precisely the Hubble law. It thus appears that we live in an expanding universe.

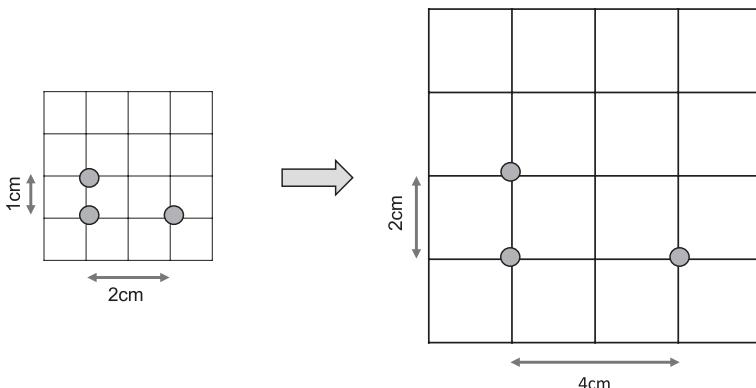


Fig. 7.4 Expanding "rubber sheet" universe. In 1 s the size doubles

7.2 A Beginning of the Universe?

The implications of Hubble's discovery were truly mind-boggling. If the distances between galaxies are getting larger, they must have been smaller at earlier times. As we follow the motion of galaxies back in time, they get closer and closer together, until they all merge at some moment of time in the past. This seems to imply that the expansion of the universe must have had a beginning. Was that the beginning of our world?

The problem of the origin of the universe, which had for centuries been the province of philosophers and theologians, had thus invaded the world of physicists and astronomers. Friedmann's models suggested that the whole universe began at a singular event a finite time ago. But for many this was too much to take. "Philosophically, the notion of a beginning of the present order of Nature is repugnant to me," wrote Sir Arthur Eddington, a prominent British astronomer. "As a scientist I simply do not believe that the Universe began with a bang." Einstein was equally disturbed. In a letter to the Dutch astronomer Willem de Sitter he wrote: "To admit such possibilities seems senseless."

And indeed, a cosmic beginning a finite time ago appeared to raise a host of perplexing problems. What actually happened at the beginning? And what caused it to happen? What determined the initial state of the universe? In the wake of Hubble's discovery, no obvious answers presented themselves. But once these problems came into focus, much of the further progress in cosmology was driven by attempts to understand the early stages of the expansion, what caused the expansion to begin, and ultimately how—and whether—the universe came into being.

7.3 The Steady State Theory

Most physicists hoped that Hubble's discovery would somehow be explained without having to postulate that the universe had a beginning. The most notorious attempt of this kind was the "steady state theory", proposed in 1948 by Fred Hoyle, Hermann Bondi and Thomas Gold, all at Cambridge University. This theory was based on the so-called "perfect cosmological principle", which asserts that the universe looks more or less the same at all times, at all places, and in all directions. An obvious implication is that the universe had no beginning in time. But how could this picture be reconciled with the fact that the universe was known to be expanding? Surely the distances between galaxies would grow and the average matter density would dilute?

To compensate for the expansion, Hoyle, Bondi and Gold proposed that matter is continuously created out of the vacuum, so that the average mat-



Fig. 7.5 Fred Hoyle (1915–2001) is best known for his contribution to the theory of stellar nucleosynthesis, explaining how heavy elements were formed in the interiors of stars. He was also the main proponent of the steady state theory and an ardent opponent of the big bang model. Yet ironically he coined the term "big bang" (in derision) during a radio broadcast for the BBC in 1949. Credit Photo by Ramsey and Muspratt, courtesy AIP Emilio Segré Visual Archives, Physics Today Collection

ter density remains constant. To achieve this, only a few atoms per cubic kilometer per century would need to materialize. So instead of one sudden creation of all matter, a very small amount of matter would need to be continuously created (Fig. 7.5).

Many physicists supported the steady state model on philosophical grounds. But ultimately, it was proven wrong. One steady state prediction was that distant galaxies, which we see as they were billions of years ago, should look more or less the same as galaxies in our neighborhood. We now know that distant galaxies are smaller, have irregular shapes and are populated by very bright, short-lived stars. Unlike nearby galaxies, many of them are powerful sources of radio waves.

The final demise of the steady state theory came with the discovery of the Cosmic Microwave Background (CMB) radiation in the mid-1960s. The detection of the CMB proved that the early universe was very hot, and the hot big bang model emerged as the standard cosmological paradigm. Cosmologists had to accept that dealing with the beginning of the universe was an unavoidable workplace hazard!

7.4 The Scale Factor

As the universe expands, the distances between all galaxies are stretched by the same factor. Similarly, if we go back in time, all distances are contracted by the same factor. The factor by which the distances change as we go from the present cosmic time t_0 to some future or past time t is called the *scale factor*; it is denoted by $a(t)$.

If two galaxies are currently separated by a distance d_0 then their separation at any other time t is

$$d = a(t)d_0. \quad (7.3)$$

At the present time t_0 , the scale factor is conventionally defined to be $a(t_0) = 1$; at earlier times $a(t) < 1$, at later times $a(t) > 1$, and at the big bang $a(t = 0) = 0$. The relative velocity of a pair of galaxies is given by the rate of change of their distance. We write $v = \dot{d}$, where an overdot is the standard physics notation for “the rate of change”. (If you are familiar with calculus, you will recognize that \dot{d} is the derivative of the distance with respect to time.) From Eq. (7.3) we find

$$v = \dot{a}(t)d_0 \quad (7.4)$$

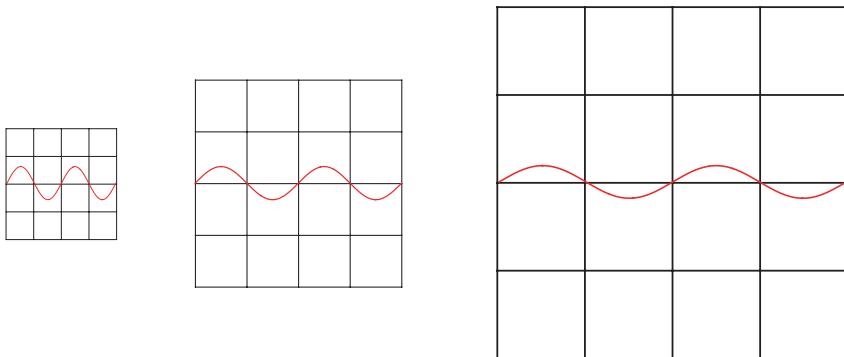


Fig. 7.6 Cosmological redshift. The wavelength of light gets stretched because space itself stretches

where \dot{a} is the rate of change of the scale factor.³ Thus, the relative velocity of the galaxies depends on how fast the scale factor changes with time. The Hubble parameter can now be found from $H = v/d$, which gives [using Eqs. (7.3) and (7.4)]

$$H = \frac{\dot{a}(t)}{a(t)} \quad (7.5)$$

Thus the Hubble parameter at any time is equal to the rate at which the scale factor is changing divided by the scale factor at that time. It is important to note that the Hubble parameter (or Hubble constant, as it is sometimes called) is constant in space, but it can change with time.

7.5 Cosmological Redshift

So far we have explained the observed redshift of light by the Doppler effect, due to the motion of galaxies away from us. We can now give an alternative interpretation, which is simpler: as the light waves travel to us, their wavelength is stretched by cosmic expansion (see Fig. 7.6).

When light leaves a distant galaxy at some time t , it starts off with a certain wavelength λ . By the time it reaches us, the universe has increased in

³If we multiply some variable by a constant, its rate of change gets multiplied by the same constant.

size by the factor $1/a(t)$, and the wavelength of light is stretched by the same factor. The wavelength λ_0 observed at present on Earth can be found from

$$\frac{\lambda_0}{\lambda} = \frac{1}{a(t)}. \quad (7.6)$$

The cosmological redshift z is defined as the fractional change in the wavelength,

$$z = \frac{\lambda_0 - \lambda}{\lambda} \quad (7.7)$$

and we have a relation between redshift and scale factor,⁴

$$z + 1 = \frac{\lambda_0}{\lambda} = \frac{1}{a(t)}. \quad (7.8)$$

Thus, by measuring the redshift of light coming from a distant galaxy, we know immediately by how much the universe has expanded since the light was emitted.

7.6 The Age of the Universe

If the universe began a finite time ago, then how old is it? To find out, we can follow the motion of galaxies back in time and evaluate how long it takes until they merge at the big bang. For a rough estimate, we shall first neglect the effect of gravity. Under this condition, any given pair of galaxies moves at a constant relative speed. Consider two galaxies at a distance d from one another. According to Hubble's law, they move apart at speed $v = H_0 d$. If they have always moved at this speed, then the time elapsed since the big bang is

$$t_0 = d/v = d/H_0 d = 1/H_0 = 4.5 \times 10^{17} \text{ s} \approx 14.4 \times 10^9 \text{ yrs.} \quad (7.9)$$

⁴For light emitted at an early epoch, when the universe was much smaller than it is today, we have $a(t) \ll 1$ and $z \gg 1$. Note that in this regime Eq. (6.1) for the Doppler shift cannot be used. It applies only to light sources moving at speeds small compared to the speed of light, that is, only to $z \ll 1$. (The symbols “ \ll ” and “ \gg ” mean “much less than” and “much greater than” respectively).

Note that this time is independent of the distance d , so it is the same for all pairs of galaxies. Cosmologists call $1/H_0$ the “Hubble time”.

We can improve this estimate, by taking into account that the universe actually has a time-varying expansion rate. Early in its history the universe decelerated due to gravity, while later on it began a period of accelerated expansion (for reasons to be discussed later). It turns out that these two effects almost cancel one another. The best modern estimates taking all the details into account give an age of 13.77 billion years. It is quite remarkable that less than 100 years ago, we did not even know that the universe contained other galaxies, yet today we can calculate the age of the universe to within a half of a percent.

7.7 The Hubble Distance and the Cosmic Horizon

Hubble's law tells us that the velocities of galaxies grow in proportion to their distance. It follows that the velocities can get arbitrarily large for galaxies sufficiently far away. This may sound alarming, since motion faster than light appears to contradict special relativity. But in fact there is no contradiction. It is important to realize that the expansion of the universe is an expansion of space, not an expansion of galaxies into some pre-existing space. The theory of relativity requires that objects cannot move past one another faster than the speed of light, but there is no limit to how fast the space between objects can expand. The distance beyond which galaxies recede faster than the speed of light is called the *Hubble distance*. We can find it by setting $v = c$ in Eq. (7.1) and solving for d ; this gives

$$d_H = \frac{c}{H_0} = 14.4 \times 10^9 \text{ ly.} \quad (7.10)$$

Another important distance scale is set by our cosmic horizon. In a universe of a finite age, there is a limit to how far we can see into space. The distance that light has traveled since the big bang is finite, and light sources that are too far away cannot be seen, simply because their light has not yet reached the Earth. We can imagine ourselves at the center of a gigantic sphere—the observable part of the universe. The boundary of this sphere is called the *particle horizon*; its radius d_{hor} is the distance to the most remote objects (“particles”) that we can possibly observe. We shall refer to the particle horizon as simply “the horizon” and will use the term “particle horizon” only

where it can be confused with another kind of horizon—the *event horizon*, which we shall later encounter.

Since the age of the universe is $t_0 \approx 14 \times 10^9$ y, you might think that the horizon distance is simply $ct_0 \approx 14 \times 10^9$ ly. You would be right if the distance from us to cosmological light sources did not change with time. But in an expanding universe a given source moves away from us while its light travels towards the Earth. Thus, by the time we detect the source, it is at a greater distance than when its light was emitted. For the most remote observable sources, the emission time is close to the big bang. The source was then much closer to us than it is now, and its present distance depends on the entire expansion history of the universe from the big bang to the present time. Calculations based on our current understanding of this history give the horizon distance

$$d_{hor} \approx 46 \times 10^9 \text{ ly}, \quad (7.11)$$

about 3 times larger than the naïve estimate. In upcoming chapters we will learn that the evolution of the early universe was first dominated by radiation, then matter and finally by “dark energy”. All of these components cause the horizon to grow in different ways (when radiation dominated, the horizon grew the slowest, and when dark energy becomes dominant, the horizon grows the fastest.). For a matter dominated universe with a flat geometry, $d_{hor}(t) = 3ct$. This gives $d_{hor} \approx 42 \times 10^9$ ly, slightly less than the actual horizon distance given in Eq. (7.11). For our purposes, an order of magnitude estimate for the horizon and Hubble distances at any cosmic time t can be found from the relation

$$d_{hor}(t) \sim d_H(t) \sim ct. \quad (7.12)$$

Light propagation and the horizon in an expanding universe are illustrated in a spacetime diagram in Fig. 7.7. The worldline of our galaxy is along the vertical axis, and the worldlines of a few other galaxies are plotted as blue curves. The galaxies get closer together as we go back in time and merge at the big bang. The slope of the curves tells us how fast the galaxies are moving away: the more steeply the curve slopes upward, the slower is the recession speed. We see from the figure that the speed grows with the distance, as required by the Hubble law. We can also tell how the recession speed changes with time for a given galaxy. The galaxies initially move apart at very high speeds. Then, as one might expect, they are slowed down by gravity. But about 5 billion years ago their motion begins to accelerate. We shall discuss the reason for this unexpected phenomenon in Chap. 9.

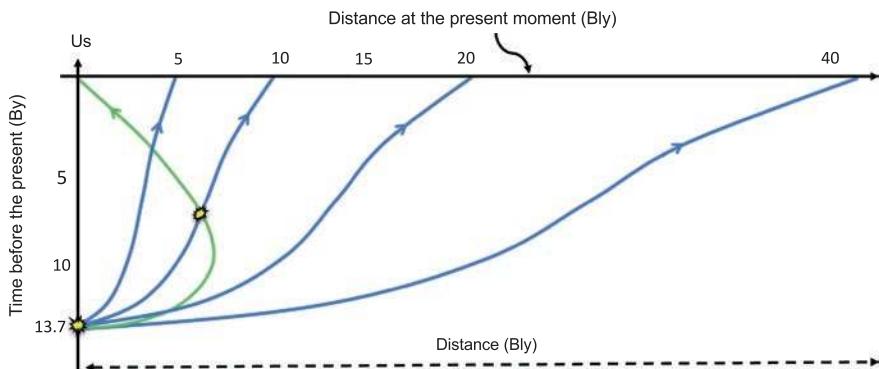


Fig. 7.7 Worldlines of galaxies (blue) and light propagation (green) in an expanding universe

Light propagation is indicated by a green line in the diagram. This line marks our past light cone. Note that it is rather different from straight-line propagation at a 45° angle that we would have in flat space. At early times, close to the big bang, light is dragged along by the expanding space, so light emitted in our direction initially moves away from us. It later turns around and, as it propagates through slower expanding space, finally approaches our galaxy along a 45° line.

Distant galaxies are now observed as they were at earlier times; these times can be found by looking at intersections of our past light cone (the green line) with worldlines of the galaxies. For example, the supernova marked in the figure occurred about 7.5 billion years ago in a galaxy that was about 7 Bly(billion light years) away at that time. The present distance to that galaxy is 10 Bly. As we look at more remote galaxies, the intersection occurs at earlier times, until we reach the galaxy whose worldline just touches our past light cone at the big bang. This galaxy is now about 46 Bly away. There are certainly more distant galaxies, but they cannot be observed, since their worldlines do not cross our past light cone. Thus, $d_{hor} \approx 46$ Bly is the cosmic horizon distance.

7.8 Not Everything is Expanding

Since the universe is expanding, you may be wondering whether or not the Solar System, the Earth, or perhaps even you yourself are expanding as well. Don't worry, you are not expanding! Objects that are bound together by forces, like atoms, planets, stars, galaxies, and even groups of galaxies, are not undergoing Hubble expansion.



Fig. 7.8 Hubble's law does not apply to galaxies bound together by gravitational forces, like the colliding galaxies in this photograph. Credit NASA, H. Ford (JHU), G. Illingworth (UCSC/LO), M.Clampin (STScI), G. Hartig (STScI), the ACS Science Team, and ESA—APOD 2004-06-12

As in the rubber sheet analogy, we can imagine some objects in an expanding universe which are fixed in space, while the space itself is being stretched by cosmic expansion. We shall refer to such objects as *comoving*. Galaxies are comoving, but only approximately: in addition to Hubble expansion, they move under the action of gravitational forces. There are many beautiful photographs of galaxies colliding (see Fig. 7.8); in fact, our Milky Way and Andromeda are falling toward one another and will collide in about 4 billion years. However, on the largest scales, when we ignore these relatively “local” motions, all matter obeys Hubble’s Law. Note that electric and magnetic fields in electromagnetic waves are not bound together by any force; that is why the light waves do get stretched.

Summary

Distant galaxies are moving away from the Milky Way, indicating that the universe is expanding. A simple relation between the speed at which a galaxy recedes from us and the distance to the galaxy was discovered by Edwin Hubble in 1929: the speed grows proportionally to the distance. This is now known as Hubble’s Law. It suggests that there is no preferred center to the expansion (all observers see galaxies receding away from their host galaxy), that the universe was much denser in the past than it is today, and that the universe had a beginning in time, the big bang, roughly 14 billion years ago.

Questions

1. State Hubble's law mathematically, and describe what it means.
2. According to Hubble's law, we should see distant galaxies receding away from us faster and faster the further out we look. Does this mean we are at the center of the expanding universe?
3. What is the universe expanding into?
4. According to special relativity the speed of light is the ultimate speed limit. Is there a limit to how fast the distances to remote galaxies can grow? Explain.
5. Does everything in the universe undergo "Hubble" expansion? For example is the distance between the Earth and Sun expanding? What about the distance between your head and toes?
6. The Andromeda galaxy is moving towards us. Does this fact falsify Hubble's Law? Explain.
7. A universe expanding according to the Hubble law, $v = Hd$, remains homogeneous and isotropic if it was homogeneous and isotropic to begin with. In such a universe, observers in any galaxy will see other galaxies receding according to the same law. Would these properties still hold if instead of the Hubble law the recession speeds of galaxies were proportional to the square of their distance, $v = Hd^2$?
8. Using Hubble's law and the nonrelativistic redshift formula $z = v/c$, calculate the distance to a galaxy that has a measured redshift of $z = 0.01$ (Assume $H_0 = 2.2 \times 10^{-18} \text{ s}^{-1}$ and $c = 3 \times 10^8 \text{ ms}^{-1}$).
9. Astronomers identified carbon lines in the spectrum of a remote galaxy and determined that their wavelengths are 1.5 times greater than the corresponding wavelengths in the carbon spectrum on Earth. By how much has the universe expanded since the time this light was emitted?
10. If the expansion of the universe has always been decelerating since the big bang, is the Hubble time greater than or less than the age of the universe? (Hint: suppose you and your friend are running a race. At some point you catch up with each other and momentarily have the same velocity. If your friend has always run with this constant velocity, and if you started out faster and have been decelerating, which one of you must have started the race first?)
11. A pulse of light is emitted from a source towards an observer, who is initially at rest with respect to the source. Consider the following two scenarios:
 - (a) After the pulse is emitted, the observer starts moving rapidly away from the source. He stops when the distance between him and the source doubles; soon after that the pulse reaches the observer. The universe does not expand in this scenario.

- (b) After the pulse is emitted, the universe starts expanding and expands by a factor of 2, so the distance between the observer and the source is stretched by the same factor. After expansion stops, the light reaches the observer. Will the observer detect any redshift in either of these situations?
12. Eternal, static models of the universe are in conflict with the second law of thermodynamics. Explain why an expanding universe can avoid this conflict.
13. Models assuming that the universe is static and infinite suffer from Olbers' paradox: each line of sight encounters a star, so the entire sky should be shining like the surface of the Sun. Explain why an expanding universe of a finite age does not have this problem.
14. The steady state theory is based on the “perfect cosmological principle” which states that on average the universe looks the same at all places, in all directions, and at all times. What observations cannot be explained by the steady state theory and why?
15. What is the cosmic horizon? If t_0 is the age of the universe, why is the horizon distance greater than ct_0 ?

8

The Fate of the Universe

Will the universe continue to expand forever, or will it eventually halt and start to collapse? We shall see that this question has a rather simple answer that depends only on the average density of the universe, ρ .¹ The larger the density, the stronger is the force of gravity that slows down the expansion. If ρ is greater than a certain critical value, ρ_c , expansion will be followed by contraction, and the universe will end in a big crunch. Otherwise, the expansion will continue eternally, and the universe will grow colder and darker as the stars exhaust their nuclear fuel, and the galaxies get further and further apart. Our goal in this chapter is to calculate the critical density ρ_c . Then, in the following chapter, we will discuss the measured value of the average density ρ and compare it to ρ_c .

8.1 The Critical Density

It is a fortunate happenstance that one does not need to employ the full blown mathematical machinery of general relativity to determine the critical density—a Newtonian analysis will lead to the correct result and will offer useful insights along the way. So let us start by considering an expanding spherical region of radius R , which represents a portion of the expanding universe. We will imagine that our hypothetical sphere is uniformly sprinkled with galaxies, which we will call “particles”. Now let us consider the

¹In this chapter we assume that there is no cosmological constant in the universe. We will revisit the fate of the universe later when we discover evidence that the cosmological constant is not zero.

motion of a “test” particle that lies on the boundary of the sphere. The gravitational effect of the rest of the sphere on this test particle is the same as if the sphere’s mass, M , were concentrated at the center. Also, the distribution of matter outside the sphere has no effect on our test particle (or any other particle within the sphere), as discussed in Chap. 2 (Fig. 8.1).

As the sphere expands, the particle (and the rest of the sphere) will be slowed down by gravity, and will either come to a halt and collapse, or will keep expanding forever. So how do we determine the outcome? We use energy conservation (this is exactly the same principle that was used when we calculated the escape speed for a projectile in Chap. 2). The particle’s energy is the sum of its kinetic and gravitational potential energy, and is given by

$$E = \frac{1}{2}mv^2 - \frac{GMm}{R} = \text{constant} \quad (8.1)$$

The mass of the test particle is m , its velocity is v , M is the mass of the whole sphere, and R is the distance of the particle to the center of the sphere. The way the sphere behaves depends on whether the total energy is negative, positive or zero.

If the total energy is negative, the particle will stop and fall back inwards. Indeed, the whole sphere will collapse. To understand why this is the case, consider the two terms that contribute to the total energy. As the particle gets further and further away, the negative potential energy term gets smaller and smaller, while the kinetic energy term is always positive. Thus, in order for the total energy to be conserved, the particle has to stop and turn around

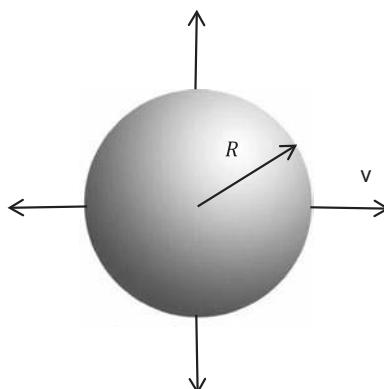


Fig. 8.1 An expanding sphere of mass M and radius R representing a portion of the universe

(if it got to infinity, it would have either zero or positive total energy depending on what the residual velocity would be at infinity). On the other hand, if the total energy is positive, then the expansion will continue, and the velocity will approach a constant value (can you determine what this value is in terms of E ?). There is a third possibility—the energy could be zero,

$$E = \frac{1}{2}mv^2 - \frac{GMm}{R} = 0 \quad (8.2)$$

This is called the critical case. The density in this case is called the critical density ρ_c .

The mass inside the sphere of radius R is related to the average mass density of the sphere via $M = \frac{4\pi}{3}R^3\rho$. Note, the sphere is expanding, so both the radius and average energy density are functions of time, while the mass remains constant. Also, from Hubble's law, the velocity of the particle is $v = HR$.

Inserting these expressions for velocity and mass into Eq. (8.2), we find $\frac{1}{2}H^2R^2 = G\frac{4\pi}{3}R^2\rho_c$, which may be rearranged to yield the expression for the critical density ρ_c in terms of the Hubble constant H and Newton's constant G :

$$\rho_c = \frac{3H^2}{8\pi G} \quad (8.3)$$

Note that ρ_c does not depend on the arbitrary radius R of the sphere. Note also that ρ_c is time-dependent because H is time-dependent.

Thus, using energy conservation, we have found that if the sphere (or the universe) has the critical density given by Eq. (8.3), expansion will continue forever, but at a speed that approaches zero (as the potential energy goes to zero, so too must the kinetic energy and hence the velocity). If the average density $\rho > \rho_c$, the expansion will halt, and will be followed by contraction and collapse. And if $\rho < \rho_c$, the universe will expand forever.

Using the current best estimate for the Hubble constant, $H_0 = 2.2 \times 10^{-18} s^{-1}$ we find $\rho_{c,0} \approx 10^{-26} \text{ kg/m}^3$ —which corresponds to only about 6 protons per cubic meter.² This is all it takes to make the universe collapse! Now, if we measure the average density ρ_0 , we should be able to forecast the ultimate fate of the universe. We will discuss more about ρ_0 in the next chapter, but before we get there, let us introduce a closely related parameter that is indispensable to cosmologists.

²The zero subscripts of H_0 and ρ_0 indicate the values of H and ρ measured at the present cosmic time.

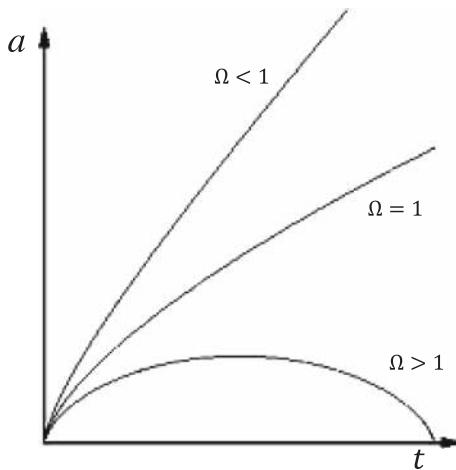


Fig. 8.2 Evolution of the scale factor (and thus the separation between any generic pair of galaxies) is determined by the density parameter. For $\Omega < 1$ galaxies approach constant recession speeds (different for different pairs of galaxies); when $\Omega = 1$ the recession speeds get smaller with time, approaching zero; and $\Omega > 1$ universes eventually contract

8.2 The Density Parameter

The density parameter is defined as the ratio of the actual (average) density to the critical density:

$$\Omega = \frac{\rho}{\rho_c} \quad (8.4)$$

We can recast the results of the previous sections of this chapter in terms of this parameter. If $\Omega > 1$, the universe eventually collapses; and if $\Omega \leq 1$, the universe expands forever (see Fig. 8.2).

Our calculation of the critical density has been performed in a Newtonian framework. Had we used general relativity, we would have found precisely the same relation between the universe's fate and the density parameter.³ In

³This is not just a lucky coincidence. Newtonian gravity is a good approximation to GR when (1) the gravitational field is weak and (2) the velocities are small compared to the speed of light. Are these conditions satisfied in our calculation? When the radius of the sphere R is sufficiently small, they are. Since R is an arbitrary parameter in our calculation, we can choose it to be small enough to ensure that the Newtonian approximation is indeed valid.

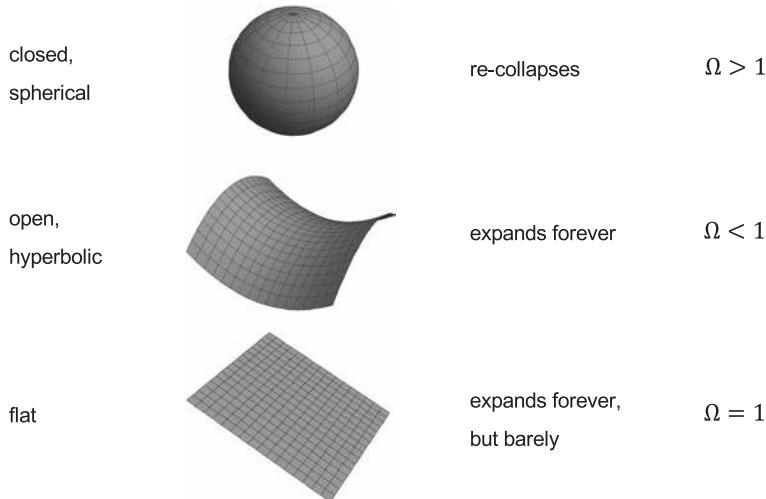


Fig. 8.3 Relation between the geometry of the universe, its fate, and the density parameter

addition, it turns out that the value of the density parameter also determines the geometry of the universe. This geometrical link can only be understood using general relativity. A closed Friedmann universe that we discussed in Chap. 5 has $\Omega > 1$; his open model has $\Omega < 1$; and a flat universe has $\Omega = 1$. The relation between the geometry of the universe, its fate, and the density parameter is summarized in Fig. 8.3.

Thus it seems as though we only have to measure Ω in order to determine the fate of the universe. However, things are not quite so straightforward—the analysis in this chapter makes use of certain assumptions regarding the contents of the universe; we will revisit the issue of our cosmic fate in the next chapter, when another important component of the universe will be introduced.

Summary

The fate of the universe is determined by its average density ρ . If ρ is larger than a certain critical value ρ_c , the force of gravity will slow the expansion down, until it halts; then the universe will contract and collapse to a big crunch. If $\rho < \rho_c$, the universe will continue expanding forever, with galaxies ultimately reaching constant recession velocities. A critical density universe with $\rho = \rho_c$ will also expand forever, but at an ever decreasing rate. Underlying these conclusions, there are certain assumptions about the contents of the universe; we revisit the issue of our cosmic fate in the next chapter.

The average density also determines the geometry of the universe: the universe is closed if $\rho > \rho_c$, open (hyperbolic) if $\rho < \rho_c$, and flat if $\rho = \rho_c$.

This relation between the average density and geometry holds regardless of the contents of the universe.

Questions

1. What properties of the universe determine whether or not it will expand forever?
2. In terms of M and R , what velocity does the test particle have in Eq. (8.2)? Can you interpret what this velocity means? If a test particle had less than this specific velocity, would it continue to move radially outwards, or would it fall back inwards?
3. Explain why gravitational potential energy is negative? Hint: Consider two objects falling towards one another from rest at a large initial distance, and think about energy conservation.
4. How is launching a projectile into space similar to the expansion of the universe?
5. At earlier cosmic times the Hubble parameter H was greater than it is now. Can you explain why?
6. Suppose astronomers living at an early epoch, when the Hubble parameter H had twice its present value H_0 , set out to determine the fate of the universe. They want to measure the density of the universe and compare it to the critical density ρ_c . Is the value of ρ_c at that epoch the same as it is now? If not, how does it differ?
7. Why is Ω such an important parameter?
8. Does the density parameter Ω change with time? If Ω is greater than one at some moment, can it become less than one at a later time?
9. Do you have a philosophical preference for a universe that ends in fire (the Big Crunch) or ice (expanding eternally)?

9

Dark Matter and Dark Energy

The composition of the heavens was a great mystery until the spectroscopic discoveries of the mid 1800s showed that the chemical elements in stars are the same as those on Earth (as we discussed in Chap. 6). But today we find ourselves grappling once again with a great mystery relating to the composition of the universe. We now have good reason to believe that most of the universe is in fact *not* made of ordinary atomic matter.

Understanding the composition of the universe is of great interest in its own right. In addition, accounting for all the matter in the universe is important for predicting its future evolution. In Chap. 8 we learned that the fate of the universe and its large-scale geometry depend on whether the density parameter, $\Omega = \frac{\rho}{\rho_c}$, is less than, equal to, or greater than one. We have already calculated the critical mass density ρ_c , so now let us turn to the measurement of the average mass density ρ . We shall see that this pursuit led to the discovery of dark matter. We will also discuss the surprising emergence of another major component of the universe called dark energy.

9.1 The Average Mass Density of the Universe and Dark Matter

On large enough scales, galaxies are approximately evenly distributed through space. Thus to calculate the *average mass density*, ρ , of the universe, a reasonable proposition seems to be to add up the masses of a large number of galaxies that span a sufficiently large volume, and then divide by that

volume. So, how does one determine the mass of each galaxy in our sample volume? One approach is based on galactic luminosity. Astronomers can use the amount and spectrum of light from a galaxy to estimate the number and types of its constituent stars. One can then add up the stellar masses, yielding a mass for the luminous matter in the galaxy. By doing so, one obtains $\Omega_{\text{stars}} \approx 0.005$. This is much smaller than unity, but we should not be too quick to conclude that the universe is open—there may still be a considerable quantity of mass hiding in stellar remnants (white dwarfs, neutron stars, and black holes), interstellar gas and dust. We need to find another way to measure the amount of mass in a galaxy that includes these “invisible” contributions.

Conveniently, we can “weigh” galaxies using our knowledge of Newtonian mechanics, just like we were able to “weigh” the Sun in Chap. 2. If a mass, M , is orbited by an object at a radius r with velocity v , then by measuring the radius and velocity, it is possible to calculate the mass inside the orbit using

$$M = \frac{v^2 r}{G} \quad (9.1)$$

Thus the Earth’s orbital radius and velocity yield the Sun’s mass.¹ Similarly, the orbital radius and velocity of a star around the center of a galaxy, yields the mass of the galaxy (that is contained within the orbit).

When astronomers plot the rotation speeds of planets (or stars) vs the distance from the center of the Solar System (or a galaxy), they obtain a “rotation curve”. The rotation curve for the Solar System is found to be precisely in accordance with the theoretical prediction: planets further out have slower orbital velocities (see Fig. 9.1). Rotation curves for many spiral galaxies have also been measured (see Fig. 9.2). However, they defy predictions. As we move from the center, the rotation velocity grows, since the orbit includes more matter. The problem arises when we get to the visible edge of the galaxy. One might expect the rotation curves to start dropping off, as they do for the Solar System. But they do not. In many cases the rotation velocity remains flat, or even increases, to distances well beyond the visible edge. This indicates that there must be a large amount of “dark matter”

¹In fact, because the planets themselves have so much less mass than the Sun, we can pick any planet, and use its distance and its orbital velocity to calculate the Sun’s mass. It doesn’t matter whether we use Mercury or Neptune, or any planet in between, we always get the same answer for the Sun’s mass.

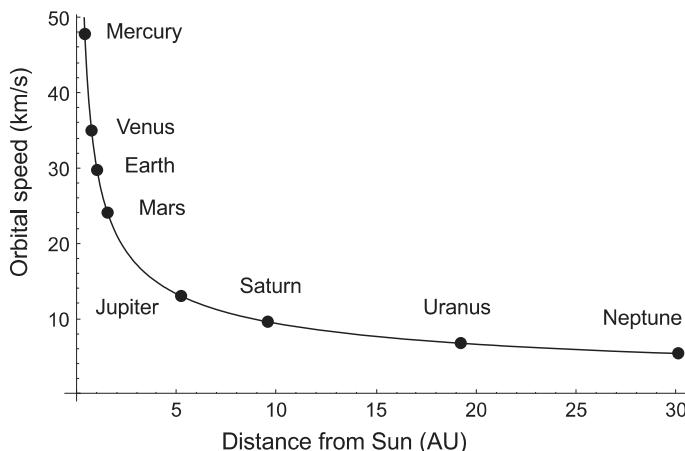


Fig. 9.1 Solar System rotation curve. The orbital velocities drop in inverse proportion to the square root of the distance from the Sun. This is known as a Keplerian fall off, and can be deduced from Eq. (9.1)

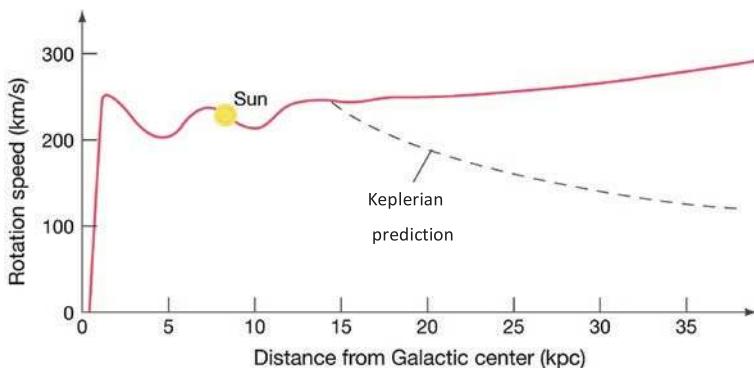


Fig. 9.2 Galaxy rotation curve. This curve can be used to determine the amount of mass lying within any given radius. The dotted curve is predicted if the mass in the galaxy ends at the visible edge of the galaxy, about 14 kpc (or 46,000 light years) from the center. The actual data do not follow the prediction, indicating that unseen mass exists beyond the visible edge. Credit Eric Chaisson [from Astronomy Today, Eric Chaisson, Stephen McMillan, Columbus (Ohio)]

beyond the visible distribution of stars. Detailed studies of rotation curves lead to the conclusion that luminous galaxies are embedded in vast dark matter halos, as illustrated in Fig. 9.3.

But how can one measure the rotation speeds for “invisible” objects beyond the visible edge? It turns out that there are vast rotating disks of hydrogen gas that extend way beyond the stars in galactic disks. The gas emits radio waves, and by measuring the Doppler shift of the radiation, the

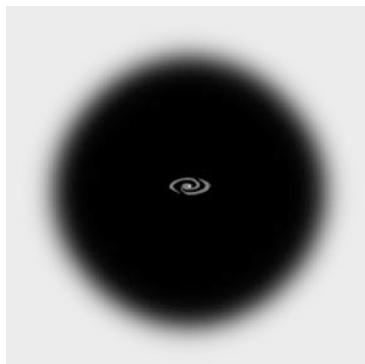


Fig. 9.3 Dark matter halo surrounding the luminous part of a galaxy

rotation curve can be extended. These measurements suggest that the dark matter halos are about 50 times more massive than the stars.

Additional evidence in support of this dramatic conclusion comes from studying *galaxy clusters*.² As we discussed in Chap. 4, light from a distant source can be gravitationally deflected by a concentration of mass that lies between the source and the observer. This can result in multiple images of the source, and in amplification and distortion of these images. If a galaxy cluster lies between us and some more distant galaxy, then the angular separation between the images of the distant galaxy on the sky and the amount by which they are distorted allows scientists to estimate how much mass is contained in the intervening cluster. Using such gravitational lensing techniques, clusters of galaxies have been weighed, and the results are consistent with those found from galactic rotation curves.

Fritz Zwicky, the Swiss born astronomer who predicted the existence of dark matter way back in the 1930s, used a different method of weighing clusters. He measured the speeds at which galaxies move in clusters and noticed that the speeds were so high that the galaxies would fly away, unless there was a large amount of unseen matter binding them to the cluster. No one took his idea seriously until four decades later, when Vera Rubin discovered, from the study of galactic rotation curves, that the universe harbors a large amount of dark matter, well in excess of the luminous matter in stars. Today, both Zwicky and Rubin are credited with the discovery of dark matter.

²A galaxy cluster is a collection of galaxies that are gravitationally bound to each other.

Another line of evidence for dark matter in galaxy clusters comes from X-ray telescopes that have revealed that clusters have very hot, tenuous atmospheres. These atmospheres are bound to their clusters by gravity, just like the Earth's atmosphere is bound to the Earth. Measurements of their temperature and X-ray radiation intensity yield two important results: (a) the atmosphere itself is several times more massive than the stars in the cluster; and (b) dark matter dominates over normal matter (atmospheric gas plus stellar mass) by a factor of about 5. Like the evidence for dark matter in galaxies, the evidence for dark matter in clusters is very strong (Figs. 9.4 and 9.5).

So what is this dark matter? As we already mentioned, part of it could be in dark stellar remnants—white dwarfs, neutron stars, or black holes. It could also include “failed stars”—low-mass objects not quite large enough to ignite nuclear reactions. But as we shall discuss in Chap. 12, none of these candidates can account for the observed amount of dark matter. We shall see that there are good reasons to believe that dark matter cannot be the usual atomic matter, but should instead consist of some exotic, as yet undiscovered particles. The particles need to be stable (so they can last for the lifetime of

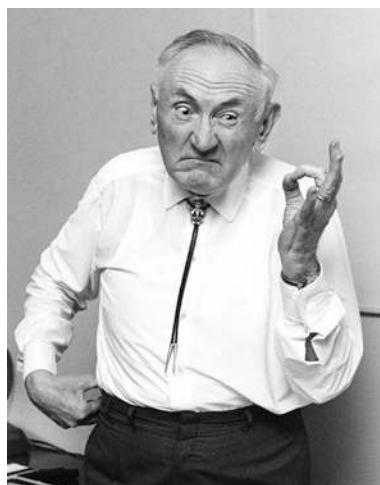


Fig. 9.4 The Swiss born Fritz Zwicky was a Professor of Astronomy at the California Institute of Technology. In addition to his discovery of dark matter, Zwicky also predicted that neutron stars would be produced in supernova explosions, and that gravitational lensing could be used to weigh galaxies and clusters. Despite the legends about Zwicky's confrontational personality (he is rumored to have called some of his colleagues “spherical bastards”, because whichever way you look at them, they are still bastards!), accounts of his compassion and generosity include how he and his wife carried out a campaign to stock libraries in war-torn Europe after World War II. Photo by Floyd Clark, courtesy of the Archives, California Institute of Technology



Fig. 9.5 During the 1970s Vera Rubin (1928–2016) found strong evidence that dark matter exists in galaxies, by studying galaxy rotation curves. In 1965 Rubin was the first female to be allowed to use the instruments at Palomar Observatory. Motivated by her own gender based professional challenges, Rubin was a strong advocate for young girls and women to pursue their scientific careers. A mother of four children, (all of whom have Ph.D.'s in the sciences), Rubin was also an observant Jew who saw no conflict between her religious views and her scientific endeavors. *Credit* AIP Emilio Segre Visual Archives, Rubin Collection

the universe) and weakly interacting (otherwise they would easily be detectable). Particle physicists have suggested a number of hypothetical candidates for dark matter particles, but for now we have to accept the fact that we do not know what most of the matter in the universe is made of.

Current measurements of the different contributions to the average density of the universe yield $\Omega_{dm} \approx 0.26$ for dark matter and $\Omega_{at} \approx 0.05$ for the atomic matter (stars and gas). The total density parameter, including both dark and atomic matter contributions, is then $\Omega_m = \Omega_{dm} + \Omega_{at} \approx 0.31$. It is less than unity, suggesting that the universe has an open hyperbolic geometry and will expand forever. This, however, is not the end of the story.

9.2 Dark Energy

We now turn to an even more mysterious ingredient of the universe, which was serendipitously discovered by two groups of astronomers in the late 1990s. The two teams, one headed by Saul Perlmutter and the other by Brian Schmidt, set out to study the expansion history of the universe, using

supernovae as standard candles. The astronomers compared the redshifts of remote galaxies to their distances, much like Hubble did, but for galaxies much further away. Redshifts were measured directly from the shift of spectral lines; and distances were determined by measuring the apparent brightness of Type 1a supernovae.

The redshift tells us how fast the galaxy was moving at the time when the light was emitted. The present velocity of the galaxy can be found from its distance. The expectation was that galactic velocities at earlier cosmic times were greater than they are today—simply because the expansion of the universe is slowed down by gravity. It therefore came as a huge surprise in 1998 when both teams discovered that galaxies are now moving *faster* than they did before.

The results of the measurements are shown in Fig. 9.6. The purple line corresponds to a universe without gravity, where galaxies are receding at constant speeds. In a decelerating universe the data points should be above this line, while in fact they are predominantly below the line. Thus the expansion of the universe is accelerating! (Figs. 9.6 and 9.7).

The redshift-distance measurements can be used to find the scale factor as a function of cosmic time, directly revealing the expansion history of the universe. The data indicate that the expansion was decelerating in the past, but in more recent times the expansion of the universe started accelerating. The turning point was roughly five billion years ago (around the time our Solar System was formed).

What could possibly cause the observed accelerated expansion? And what could explain the transition from deceleration to acceleration? It is as though attractive gravity suddenly flipped to become repulsive.

We have already encountered one instance where gravity can be repulsive: remember the vacuum energy density, or the cosmological constant? Let's suppose the vacuum has a nonzero mass density ρ_v . This will produce a repulsive force. If ρ_v is sufficiently large, this force will overcome the attractive gravity of matter. The result will be an accelerated expansion of the universe.

The vacuum density ρ_v remains constant in time, while the density of matter ρ_m changes with the expansion of the universe. The volume of any given region grows like the cube of the scale factor,

$$V(t) \propto a^3(t) \quad (9.2)$$

If M is the mass of matter contained in the region, then the density of matter is

$$\rho_m(t) = \frac{M}{V(t)} \propto \frac{1}{a^3(t)} \quad (9.3)$$

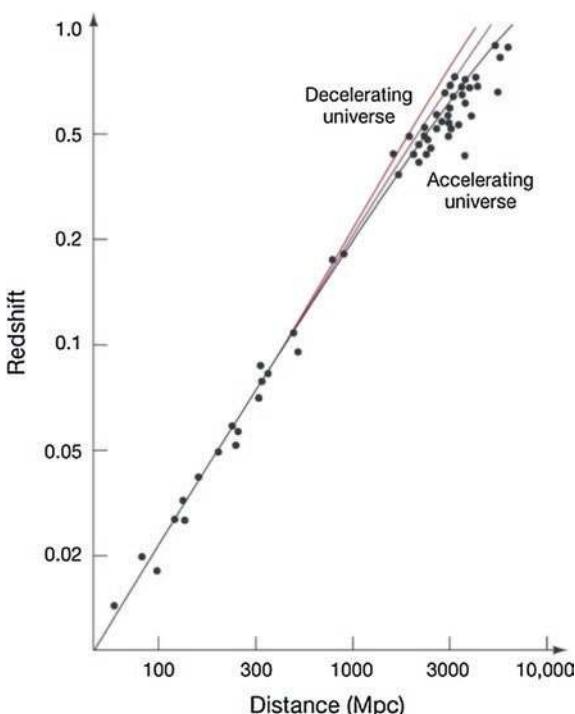


Fig. 9.6 Redshift-distance graph for distant supernovae. The *red* (*upper*) and *black* (*lower*) lines correspond to decelerating and accelerating universes, respectively. The *purple line* (*in the middle*) is for a universe without gravity, where galaxies move at constant speeds. The data indicate that our universe is accelerating. Credit Eric Chaisson [from Astronomy Today, Eric Chaisson, Stephen McMillan, Columbus (Ohio)]



Fig. 9.7 Saul Perlmutter of the Lawrence Berkeley Laboratory, Brian Schmidt of Mt. Stromlo Observatory in Australia and Adam Riess of Johns Hopkins University won the 2011 Nobel prize in physics for their role in the discovery of accelerated expansion of the universe. Credit (c) Nobel Media AB photo Ulla Montan

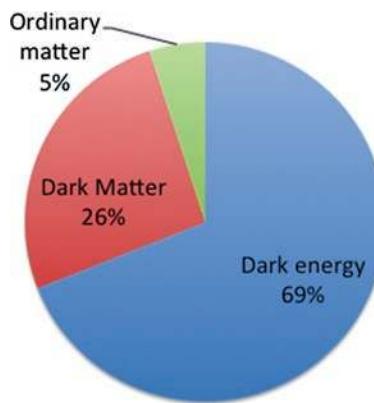


Fig. 9.8 The present composition of the universe

So at early times, when the scale factor is very small, ρ_m must be much greater than ρ_v . Thus the attractive gravity of matter overwhelms the repulsive gravity of the vacuum. But, as the universe expands, the matter density is diluted, and eventually it drops below the vacuum density. At that point the cosmic acceleration begins. [More precisely, acceleration begins when $\rho_v > \rho_m/2$; see Chap. 5, Eq. (5.3)].

If ρ_v is indeed the reason for the accelerated expansion, we need to determine how large it is. The best fit to the data is obtained for $\rho_v \approx 2.2 \rho_{m0}$, where ρ_{m0} is the present matter density.³ The corresponding vacuum density parameter is

$$\Omega_{vac} = \rho_v / \rho_c \approx 0.69 \quad (9.4)$$

The vacuum energy, which is often called “dark energy”, is thus the dominant component of the universe (see Fig. 9.8). An intriguing consequence of Eq. (9.4) is that the total density parameter is very close to unity: $\Omega_{tot} = \Omega_{vac} + \Omega_m \approx 1$. It seems as though the universe is perched on the borderline between being open and closed—it is flat, or at least very nearly flat.⁴

³Note that ρ_{m0} includes both dark and atomic matter.

⁴Several other independent measurements also indicate that the universe is flat, further adding confidence to the notion that the universe is filled with dark energy. We will address these findings in later chapters.

9.3 The Fate of the Universe—Again

Dark energy also has important implications for the future of the universe. The Friedmann relation between the density parameter and the fate of the universe that we discussed in Chap. 8 only holds when there is no vacuum energy. (We emphasize that the relation between Ω and the *geometry* of the universe is always valid). Once the universe is dominated by vacuum energy (as it is today), it will keep expanding forever, due to repulsive gravity, regardless of the value of Ω . The universe will double in size about every 10 billion years. The velocities of the galaxies will also double on the same time scale.

This kind of expansion is called exponential. As the recession speeds of galaxies exceed the speed of light, they will leave our observable region, never to be seen again. The universe will thus get emptier and emptier, until (in a few trillion years), no stars outside our local galaxy cluster will be visible at all. As for the stars themselves, eventually they will exhaust their nuclear supplies, and their embers will fade into the frigid blackness of space. We should rejoice that we live at the current cosmic epoch, under a bejeweled sky brimming with clues about our cosmic origins.

We need to take a moment to reflect on a very curious feature of our universe. The vacuum energy density was much smaller than that of matter in the early universe, and it will get much greater than the matter density in the future. We happen to live at a very special epoch when these two densities are comparable: $\rho_{vac} \approx 2\rho_{m0}$. Do you think this is simply a coincidence?

Summary

The mass density in luminous stars amounts to a small fraction of the critical density. By studying the rotation curves of galaxies, as well as gravitational lensing and galaxy velocities in clusters of galaxies, we find that, by mass, there is roughly 5 times more dark matter, than all the matter in stars and gas. We don't know what the dark matter is made of, but we now know it is not composed of protons, neutrons and electrons, like ordinary matter.

Even with dark matter included, the total matter density is still less than the critical value. But recent observations of supernova explosions have revealed that apart from dark matter, the universe contains yet another mysterious dark component. Observations indicate that the expansion of the universe is now accelerating with time. The most likely cause of this acceleration is the existence of a space-filling vacuum energy, called “dark energy”, which has a repulsive gravitational effect. The discovery of dark energy

changes the fate of the universe—it will continue to expand, regardless of whether it is open or closed.

Questions

1. Astronomers have found that $\Omega_{\text{stars}} \approx 0.004$. Briefly explain how they made this measurement.
2. If there is matter in and around galaxies that is not luminous, how can we know it's there?
3. Describe two methods astronomers use to infer that there are large amounts of dark matter in galaxies and clusters of galaxies.
4. Would an electrically charged particle be a good dark matter candidate? Why?
5. Consider a galaxy containing stars that are concentrated within a radius $R_s = 25,000$ light years from the galactic center. At this radius, the stars are observed to be rotating at 200 km/s around the galactic center. Find the total mass of matter contained within the radius R_s . Assuming that most of this mass comes from stars with mass comparable to the Solar mass, estimate how many stars are in this galaxy.
If the same velocity is measured from hydrogen gas at 75,000 light years, how much mass is contained in this larger orbit?
6. Do you find it upsetting that normal atomic matter is just a small percentage of the total matter content in the universe?
7. How do you think Einstein would have reacted to the 1998 discovery that the cosmological constant may not be zero?
8. Today the universe is undergoing accelerated expansion. If we go back in time, was there ever a period when the expansion of the universe slowed down?
9. Roughly when did the universe start to accelerate its expansion—in the last quarter century, several thousand years ago, or a few billion years ago?
10. In the early universe $\rho_v \ll \rho_m$, but today $\rho_v \approx 2\rho_m$. Briefly explain how the vacuum energy density came to dominate the matter density, by considering how ρ_v and ρ_m do or don't change with time.
11. How is dark matter different from dark energy?
12. Show [using Eq. (5.2)] that any “stuff” with a negative pressure that has an absolute value $|P| > \rho/3$ would be gravitationally repulsive.
13. Use Eqs. (5.1) and (5.2) to show that expansion is accelerated when $\rho_v > \rho_m/2$.

10

The Quantum World

Modern physics began with two revolutions at the turn of the 20th century. The first revolution, which radically changed our concepts of space and time, was single handedly accomplished by Einstein with his special and general theories of relativity. The development of quantum mechanics by a number of physicists ushered in the second revolution, which shook the foundations of physics even more than the first. Quantum mechanics was developed as a theory of the microworld but as we shall see, quantum effects are essential in the early universe and even play a role on the largest cosmic scales.

10.1 Quantum Discreteness

According to quantum mechanics, at the microscopic level electromagnetic waves consist of *photons*—small bundles (or quanta) of electromagnetic energy. Photons always travel at the speed of light and have zero rest mass. The energy of a photon is inversely proportional to its wavelength λ , and is given by

$$E = \frac{hc}{\lambda}, \quad (10.1)$$

where h is Planck's constant: in SI units $h = 6.6 \times 10^{-34} \text{ J s}$. Scientists often use the reduced Planck constant, $\hbar \equiv h/2\pi$ —we will use both. Because h is

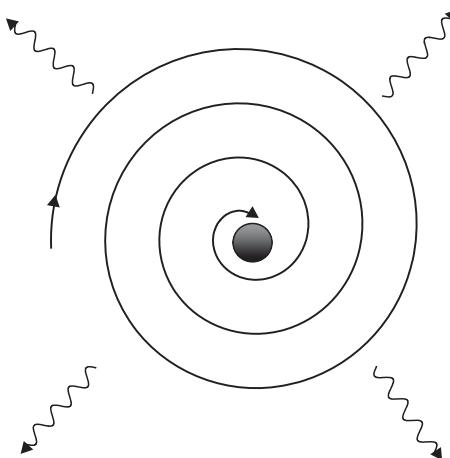


Fig. 10.1 Classical mechanics and electromagnetism predict that orbiting electrons will radiate electromagnetic waves, lose energy and spiral in toward the nucleus

such a tiny number, a photon typically carries a tiny amount of energy. The classical wave description of light is accurate when we have a large number of photons; for example, a 100 W light bulb produces about 10^{19} photons per second.

Quantum discreteness is also manifested in atomic structure. The early 20th century “planetary” model of the atom consisted of negatively charged electrons orbiting around a positively charged nucleus, much like the planets orbit the Sun. However, Maxwell’s theory of electromagnetism predicts that charged particles moving along curved trajectories will radiate electromagnetic waves. Thus physicists were puzzled by how the electrons could maintain their stable orbits, and avoid continuously radiating away energy that would cause them to spiral into the nucleus (Fig. 10.1).

In quantum theory, atomic electrons are allowed to occupy only a discrete set of orbits, with each orbit having a specific energy. An electron can emit a photon and jump to a lower orbit, as shown schematically in Fig. 10.2.¹ This process must conserve energy, so the energy of the photon must be equal to the energy difference between the two orbits. The inverse process is also possible—an electron can absorb a photon and jump to a higher-energy orbit. Thus atoms can emit and absorb photons of only specific energies (and wavelengths). The existence of discrete energy levels is essential for spectroscopic measurements, which provide much of the information that we have about the universe.

¹Electron orbits are actually somewhat fuzzy and are more accurately described by wave functions; see Sect. 10.3.

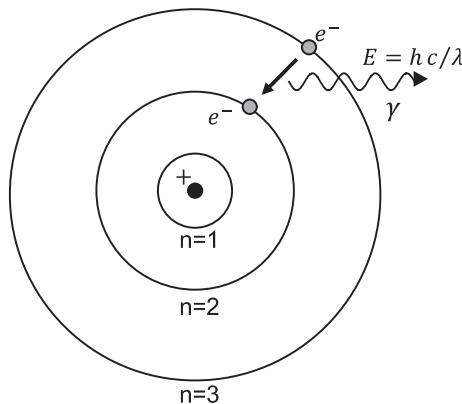


Fig. 10.2 Quantum model of the atom. Larger orbits have higher energy. $n = 1$ is the lowest energy level of the atom. Here, a photon is emitted when an electron jumps from the $n = 3$ level to the $n = 2$ level. The energy of the photon is equal to the energy difference between the levels, so the total energy is conserved

10.2 Quantum Indeterminism

The quantum world is fundamentally unpredictable. We can never know for certain where a given particle will be or how fast it will move; the best we can do is to predict probabilities for possible future positions and velocities. This is in contrast to classical, Newtonian physics, where the entire future history of a particle can be predicted from its position and velocity at some initial moment (Fig. 10.3).

At the core of quantum physics is the uncertainty principle, discovered by Werner Heisenberg in 1927. It states that the position and velocity of a particle cannot be simultaneously determined. The more precisely we measure the position, the greater is the uncertainty in the velocity, and vice versa. This is encoded in the equation,

$$\Delta x \cdot \Delta v > \frac{\hbar}{4\pi m} \quad (10.2)$$

where Δx and Δv are respectively the uncertainties in the particle's position and velocity, and m is the particle's mass (this equation applies only to non-relativistic particles). If we make Δx very small, then Δv will get large—in a sense, the more we try to localize the particle, the more it tries to “escape”. A quantum particle is thus inherently fuzzy and cannot be assigned a definite trajectory.

Macroscopic objects, like planets or billiard balls, follow their classical trajectories with very high probability, which is why the motion of planets can



Fig. 10.3 Werner Heisenberg. Credit AIP Emilio Segre Visual Archives, Gift of Jost Lemmerich

be predicted for many centuries to come. But for small particles, like electrons, deviations from classical motion can be very large. Such unpredictable deviations are called quantum fluctuations.

One of the most striking examples of quantum fluctuations is illustrated in Fig. 10.4. A ball lies at a low point in a one-dimensional landscape, separated by a hill from a still lower valley. In the world of classical physics, the ball would stay where it is, unless someone kicks it, providing the energy necessary to get over the hill. But in the quantum world, there is a non-zero probability that the ball will suddenly and discontinuously emerge on the other side of the hill and start rolling down. This process is called “quantum tunneling”. The larger the energy barrier (or height and/or width of the hill) that needs to be surmounted, the smaller is the tunneling probability.

While tunneling might sound like an exotic quantum effect, it has many real world consequences and applications. It explains, for example, the phenomenon of alpha radioactivity, where an alpha particle (consisting of two

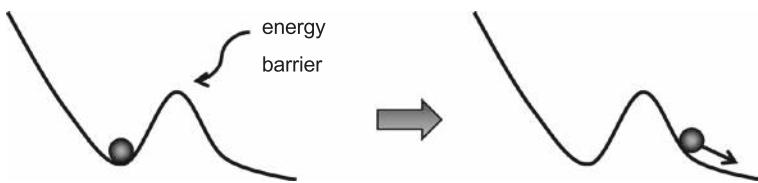


Fig. 10.4 Quantum tunneling through an energy barrier



Fig. 10.5 Erwin Schrodinger. Credit Photograph by Francis Simon, courtesy AIP Emilio Segré Visual Archives

protons and two neutrons) is emitted from inside a nucleus, despite the energy barrier produced by attractive nuclear forces. Also, the scanning tunneling electron microscope can be used to see individual atoms on the surface of a material. When a sharp conducting probe is scanned above a surface, the distance between the probe and the surface will vary slightly, depending on the arrangement of the surface atoms. When the surface atoms are closer to the probe it will be easier for electrons to tunnel from the surface to the probe, which then registers a current. Thus, by measuring the rate at which electrons tunnel from the surface to the probe, one can image the individual bumps and depressions of atoms on the surface (Fig. 10.5).

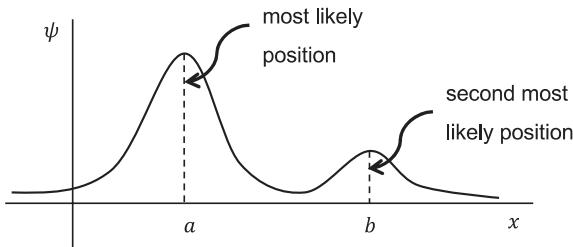


Fig. 10.6 Wave function of a particle

10.3 The Wave Function

In quantum theory, a particle is mathematically described by a wave function $\psi(x, t)$, which is a function of position x and time t . It contains all the information we can have about the particle. The shape and time evolution of the wave functions are determined by the so-called Schrodinger equation, derived by Erwin Schrodinger in 1927. The wave function does not tell us where the particle is located; it only determines the *probability* to find it in one location or another.² Suppose at some moment of time an electron is described by the wave function shown in Fig. 10.6. If we measure the position of the electron, we are most likely to find it near position a , where ψ has the largest magnitude. The second most likely possibility is to find it near position b , and there is some non-zero probability that it is at any other location where the wave function is not zero.

If we perform many identical measurements, their outcomes will be distributed according to the probabilities predicted by the wave function.

Prior to a measurement, an electron described by the wave function in Fig. 10.6 does not have any definite position. We say it is in a superposition of states corresponding to different positions. Once we perform a measurement, we know where the electron is at that moment, so the wave function “collapses” to a peaked shape around that point, as in Fig. 10.7. The electron will not generally stay localized, and the peak of the wave function will immediately start to spread. Once again, its time evolution can be found by solving the Schrodinger equation, and the resulting wave function can be used to determine the probabilities of future measurements.

If we prepare a large number of electrons in the same quantum state (described by the same wave function) and perform identical experiments

²More precisely, the probability distribution is given by the square of the wave function.

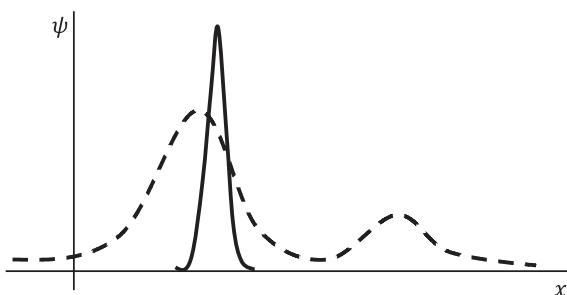


Fig. 10.7 Collapse of the wave function. The *solid line* represents the wave function after measurement

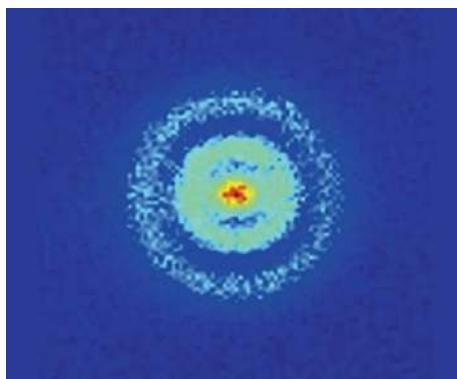


Fig. 10.8 The measured probability distribution for one of the energy levels in a hydrogen atom. Each *dot* on this image represents the measured location of an electron relative to the nucleus. By using a large collection of hydrogen atoms, whose electrons are all in the same energy level, this image represents the probability distribution, or wave function, of the electron. Credit Stodolna et al. *Phys. Rev. Lett.* 110, 213001

measuring the positions of the electrons, the data points will provide an image of the probability distribution, as shown in Fig. 10.8.

The wave function description is not limited to the positions of particles; it can be applied to any quantum system. As another example, consider a radioactive atom, whose nucleus can decay by emitting an alpha particle. Radioactive decay is a fundamentally random process, so you cannot predict the time of decay. You can only determine the decay probability per unit time (say, per hour). Suppose you checked that the atom is intact at some initial moment and placed it in a sealed box, so you cannot observe it. Then, at a later time, the wave function of the atom will be a superposition of decayed and un-decayed states. Just like the electron in the previous example

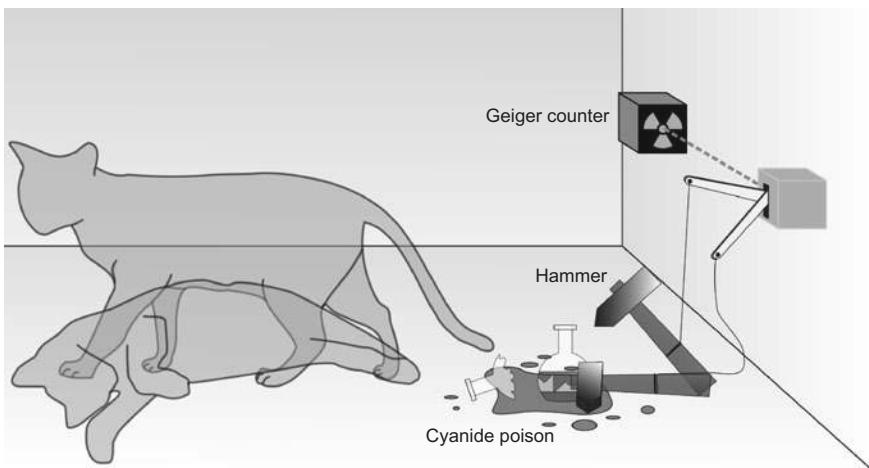


Fig. 10.9 Schrodinger's cat. Credit Dhatfield, wikipedia

did not have any definite position, the atom has no definite state of decay. You can open the box and inspect the atom; then the wave function will collapse to either a decayed or un-decayed state, with probabilities that you can calculate. It appears that the atom “makes up its mind” at the last moment, when the measurement is performed.

To highlight just how bizarre this is, Erwin Schrodinger proposed the following thought experiment. Imagine there is a cat in a perfectly sealed box containing a radioactive atom and a Geiger counter. There is also a flask of cyanide poison in the box. If the radioactive atom decays, the Geiger counter detects a signal that triggers a hammer to smash open the poison, instantaneously killing the cat. We should now describe the entire content of the box by a wave function, and it will be a superposition of two states—an intact atom plus a living cat and a decayed atom plus a dead cat. The cat is thus in a superposition of “dead” and “alive” states! If we were to open the box and look inside, the cat would suddenly become either “alive” or “dead”—its wave function would “collapse” (Fig. 10.9).

If you are scratching your head, rest assured you are not alone. The probabilistic interpretation of the wave function that we outlined above was developed by Max Born and by Niels Bohr and his colleagues at his institute in Copenhagen; it is called the Copenhagen interpretation. But some of the founders of quantum mechanics never accepted quantum indeterminism. Most notable amongst them was Einstein who quipped: “God does not play dice with the universe”.

10.4 Many Worlds Interpretation

In 1957 a Princeton graduate student, Hugh Everett, proposed an alternative interpretation of quantum mechanics which postulates that the wave function never collapses. Instead, all possible outcomes of any measurement do occur, but they occur in “parallel” universes, which have no contact with one another.

With every measurement of a particle’s position, the universe branches into multiple copies of itself, where the particle is found to be in all possible places.

The branching process is described by the Schrodinger equation and is fully deterministic. But we cannot predict which of the parallel universes we will find ourselves in, and thus the outcomes of *our* measurements can still be determined only probabilistically. Everett showed that the probabilities come out exactly the same as when one uses the Copenhagen interpretation. Everett’s approach is now called “the many worlds interpretation” (Fig. 10.10).

Debate about the meaning of the wave function still continues. But despite this uncertainty about its philosophical foundations, quantum mechanics is a tremendously successful theory which has been crucial for



Fig. 10.10 Hugh Everett circa 1964. Credit Courtesy of Mark Everett. Hugh Everett III Manuscript Archive, UCISpace@the Libraries Permanent url: <http://hdl.handle.net/10575/1060>

our understanding of atomic structure, chemistry, biochemistry, particle physics and so on. All of its predictions have been borne out by experiments with incredible precision. It is also the theoretical framework that underpins the technology of transistors, atomic clocks, lasers, superconductivity, etc.

Since the choice of interpretation does not affect any predictions of the theory, most physicists simply disregard the philosophical problems and follow the dictum “*Shut up and calculate!*” This attitude works fine, except in cosmology, where one might want to apply quantum theory to the entire universe. The Copenhagen interpretation, which requires an external observer to perform measurements on the system, cannot even be formulated in this case: there are no observers external to the universe. Cosmologists, therefore, tend to favor the many worlds interpretation.

Summary

The physics of the microworld is governed by the inherently discrete quantum mechanics. In particular, the classical picture of electromagnetic waves gives way to a quantum description in which light consists of photons that carry discrete amounts of energy. Atomic electrons also have quantized energies that can increase or decrease only by discrete amounts via the absorption or emission of photons. This gives rise to the spectroscopic absorption and emission lines.

In contrast to the classical, deterministic universe, the quantum world is fundamentally unpredictable. Even if we have complete information about a quantum system, we can only make probabilistic predictions about its future evolution. Macroscopic bodies, like cars or tennis balls, behave nearly classically, but in the microworld, unpredictable deviations from the classical motion, called quantum fluctuations, are typically large.

In quantum physics a particle is described by a *wave function*, which determines the probability for the particle to be in various locations. According to the Copenhagen interpretation of quantum mechanics, once we perform a measurement, the wave function “collapses” and the particle then momentarily has the measured position. An alternative interpretation of quantum mechanics, called “the many worlds interpretation”, asserts that all possible outcomes of the measurement occur in disconnected “parallel” universes. We cannot determine which universe we are in, thus the future events we expect to observe can only be predicted probabilistically. Regardless of how we interpret quantum mechanics, its predictions remain the same.

Questions

1. Einstein was one of the founders of quantum mechanics. Yet he still felt uneasy about the probabilistic nature of quantum mechanics. Do you find the probabilistic world of quantum mechanics more or less appealing than the deterministic classical universe?
2. Epicurus asserted that atoms move deterministically, but occasionally experience random “swerves”. He thought the swerves were necessary to explain the existence of free will. Quantum mechanics appears to provide something very similar to the swerves. Do you think this helps to explain free will?
3. Can you think of examples where it is better to think of light as a wave, and light as a particle?
4. Can you explain why atoms have specific absorption and emission spectral lines?
5. Do you think that with improved technology we will be able to overcome Heisenberg’s uncertainty principle, and measure the exact position and velocity of an electron?
6. Could a stationary grape in a glass bowl spontaneously appear outside the bowl? Compare the classical and quantum mechanical “answers”.
7. What is a quantum fluctuation?
8. What do physicists mean when they talk about the “collapse of the wave function”?
9. Discuss and compare the Copenhagen and many-worlds interpretations of quantum mechanics. Which one do you prefer?
10. Do you think the many worlds interpretation can ever be disproved?

11

The Hot Big Bang

In an expanding universe, matter is diluted as the volume of the universe gets larger. Conversely, if we follow the expansion backwards in time, we find that the universe was denser in its past than it is today. In fact, the density of the universe grows without bound as we wind the clock back to the big bang. Furthermore, the temperature of the universe also soars to extremely high values. How do we know this? And what does this imply about the conditions of the early universe?

11.1 Following the Expansion Backwards in Time

The idea of a *hot* big bang was conceived by the Russian-born physicist George Gamow in the late 1940's (Fig. 11.1). It was based on the simple observation that gases cool down when they expand and conversely heat up when compressed. The temperature of a gas is a measure of the average kinetic energy of its constituent particles. The faster the particles move, the higher the temperature. So let us consider the energetics of particles bouncing off the walls in a box (see Fig. 11.2). When the wall is stationary, any given particle will bounce off at the same speed as it hits the wall. There will be no loss of kinetic energy. However, if the wall is retracting away from the particle, then the particle will rebound at a lower speed. Thus, in an expanding box, every time a particle collides with a retracting wall, it will lose kinetic energy. This loss of energy manifests itself as a decrease in temperature.



Fig. 11.1 George Gamow's (1904–1968) many significant contributions to physics include being the first to understand radioactivity in terms of quantum mechanics and laying the groundwork for the hot big bang cosmology. He was a great popularizer of science and was known for his risqué sense of humor. In 1933 he defected from the Soviet Union, and he moved to the United States in 1934. Credit AIP Emilio Segrè Visual Archives, George Gamow Collection

The same cooling effect occurs in an expanding universe, even in the absence of walls. To understand this, let us consider how velocities of gas particles change as they travel through the universe. Suppose a particle flies by galaxy A at velocity v and moves on towards some distant galaxy B. Galaxy B is itself moving away from A at velocity u , determined by Hubble's

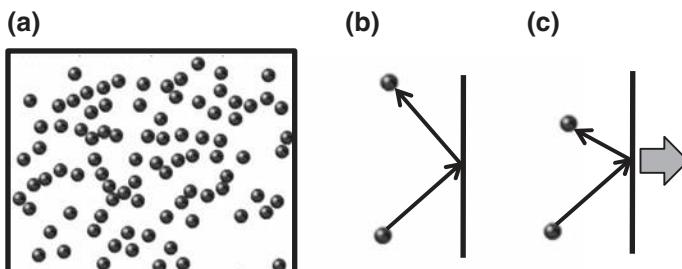


Fig. 11.2 **a** Particles in a box. **b** Particle bouncing off a stationary wall. **c** Particle bouncing off a moving wall. It rebounds with a lower speed than in **(b)**

law. So when the particle catches up with galaxy B, the observers in B will see it moving at a reduced speed, $v - u$. Galaxy C, which is at a greater distance from A, is moving away at a greater speed, so when the particle eventually catches up with C, its observed speed will be further reduced. This applies to all particles and all observers in an expanding universe. As time goes on, observers will see the particles moving slower and slower—which means that any gas filling the universe will be cooling down.

Conversely, if we follow the universe backwards in time, it will get hotter and hotter. As we will see later, the temperature in the early universe is inversely proportional to the scale factor, $T \propto 1/a$. Thus the universe apparently becomes infinitely hot as the scale factor approaches zero at the big bang. What happens to the matter content of the universe under these extreme conditions?

Everything around us consists of molecules that are composed of different types of atoms, held together by chemical bonds. Each atom is made up of electrons swirling around nuclei, which in turn consist of protons and neutrons. None of these components of matter could have existed at the early moments of the nascent universe. They would have been destroyed as energetic particles smashed into one another at super-high temperatures.

The chemical bonds that hold atoms together in molecules break at about 500 K¹; atoms break up into nuclei and electrons at roughly 3000 K; and nuclei split into protons and neutrons at approximately 10^8 K. At still higher temperatures, above 10^{12} K, neutrons and protons (collectively known as nucleons) break up into their elementary constituents, called *quarks*. All

¹One degree Kelvin is equal to one degree Celsius. The Kelvin scale however starts at absolute zero (the lowest possible temperature), which is -273.15°C . For very high temperatures close to the big bang, there is not much difference between the two scales.

complex structures disintegrate as temperature increases. Consequently, the physical state of matter in the early universe was much simpler than it is today. It was just a hot and dense mixture of subatomic particles, which is often called “the primeval fireball”.

As we refine our understanding of the fireball, we shall discover that, in addition to the particles that make up atoms, it included other particle species, such as the weakly interacting neutrinos. But most importantly, the fireball was pervaded by intense electromagnetic radiation, as we shall now discuss.

11.2 Thermal Radiation

Let us first recall that at the microscopic level electromagnetic waves consist of photons. Important things to remember about photons are that their energy is inversely proportional to their wavelength,

$$E = h \frac{c}{\lambda}, \quad (11.1)$$

and that they can be emitted and absorbed by electrically charged particles. Figure 11.3a illustrates a collision of two particles, which is accompanied by the emission of two photons. In Fig. 11.3b a charged particle absorbs a photon and then emits another one. In the super-dense early universe, these emission and absorption processes occur at a fierce rate, and equilibrium is quickly established where photons are mixed with other particles and are emitted at the same rate as they are absorbed. From the macroscopic point of view, this gas of photons can be pictured as electromagnetic radiation consisting of waves with different wavelengths.

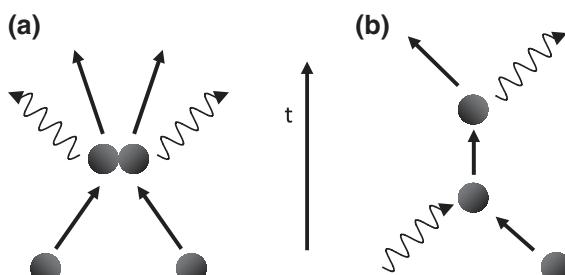


Fig. 11.3 a Photons emitted by colliding charged particles. b A charged particle absorbs a photon and then later emits another one

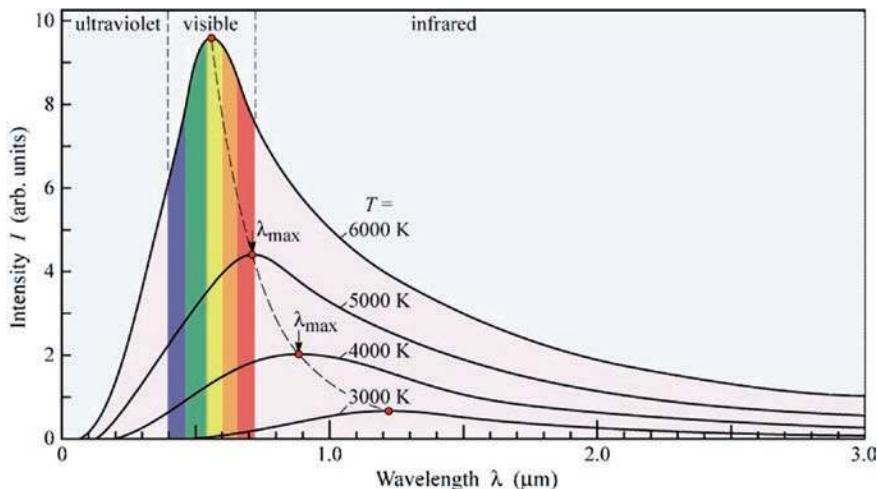


Fig. 11.4 The spectrum of thermal radiation at various temperatures. The color bands correspond to the wavelengths of visible light. λ_{\max} is the wavelength corresponding to the maximum intensity for a given temperature. Credit "Physical Foundations of Solid State Devices", by E. Fred Schubert (EFSchubert@rpi.edu), 275 pages, 2015, available at the [Google Play Store for US\\$ 8.00](#) (ISBN-13: 978-0-9863826-2-8)

Electromagnetic radiation that is in equilibrium with matter at some temperature is called thermal radiation. The higher the temperature, the higher the intensity (or the energy density) of the radiation. Quantitatively, the total intensity is proportional to the 4th power of the temperature,

$$\rho \propto T^4. \quad (11.2)$$

This intensity is spread over a range of wavelengths, with a distribution (or spectrum) that depends only on the temperature; it is shown in Fig. 11.4 for several different temperatures, and is called a thermal spectrum. The form of this distribution was derived by the German physicist Max Planck at the turn of the 20th century (Fig. 11.5).

The peak intensity occurs at a wavelength inversely proportional to the temperature,

$$\lambda_{peak} \propto 1/T. \quad (11.3)$$

Most of the photons in thermal radiation have wavelengths around λ_{peak} , and it follows from Eq. (11.1) that the typical energy of photons grows in proportion to the temperature,

$$E \propto T. \quad (11.4)$$

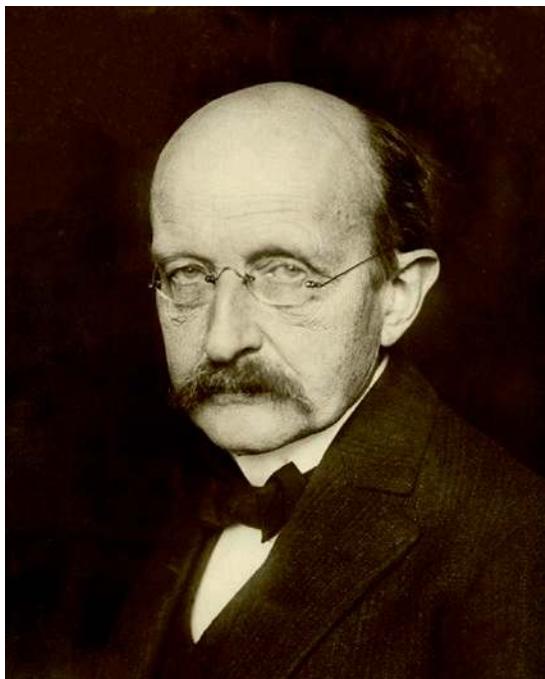


Fig. 11.5 Max Planck derived a formula for the spectrum of thermal radiation in 1901, laying the foundation for quantum mechanics

Any macroscopic object at a non-zero temperature emits radiation with an approximately thermal spectrum. The details of the spectrum depend on the material of the object—specifically on how it absorbs and reflects electromagnetic waves. The spectrum is exactly thermal only for an ideal black body, which absorbs all incident radiation.² The thermal spectrum is therefore sometimes called the black body spectrum.

²This is not difficult to understand from the following thought experiment. Consider a black body in equilibrium with thermal radiation at some temperature T . The black body absorbs all incident radiation, and in order to maintain equilibrium, it has to emit radiation at the same rate and with the same spectrum.

An ideal black body at zero temperature would look black even if you were to shine light on it. The reason is that it does not reflect any incident light. But at non-zero temperatures “black bodies” are not really black, since they glow with thermal radiation. Stars are good examples of almost ideal black bodies. The surface temperature of the Sun is 6000 K, and the corresponding peak wavelength is right in the middle of the visible spectrum.³ From Eqs. (11.2) and (11.3) we can tell that a star with a surface twice as hot as the Sun would have a total intensity that is 16 times higher, and a peak wavelength that is half as much. At human body temperature (about 300 K), the peak of thermal radiation is in the infrared range, so humans and animals all glow in the infrared. At the extreme temperatures of the premeval fireball, shortly after the big bang, the photon energies were much higher and their wavelengths much shorter than those of visible light.

11.3 The Hot Big Bang Model

The starting point of the hot big bang model is an expanding fireball of elementary particles and photons. Assuming that the universe was homogeneous and isotropic, the fireball uniformly filled the entire space. One of the main goals of cosmology is to explain how the universe evolved from this simple state to what it is today.

As the universe expands, the fireball dilutes and cools down, and complex structures begin to form. When the universe is roughly a minute old, the temperature T drops to 10^9 K, and protons and neutrons start to combine to form atomic nuclei. This is called *nucleosynthesis*, and will be discussed in Chap. 13. By the time the universe is about 380,000 years old, the temperature cools to $T = 3000$ K, and electrons combine with nuclei to form neutral atoms. This process is called “recombination”. Eventually stars, galaxies, and galaxy clusters are pulled together by gravity.

Today we find ourselves having front row seats from which to view this history: *as we look further out into the universe, we also look back in time*. If we look at a supernova 10 billion light years away, we see it as it was 7.5 billion years ago (see Sect. 7.7). If we look far enough, we will see the universe as it was when galaxies and the first stars were being formed. What if we look still

³Solar radiation has comparable intensity at all wavelengths in the visible spectrum. This should be perceived as white light, and indeed the Sun looks white when viewed from outer space. However, to observers on Earth, the Sun often looks yellow. This is mostly because the blue part of the spectrum is scattered by the Earth’s atmosphere.

further, beyond galaxies, as far as our telescopes can reach? We will see the primordial fireball. It is there, in all directions on the sky.

Unfortunately, we cannot see all the way back to the big bang. At very early times the Universe was opaque because photons were frequently scattered by charged electrons and nuclei. However, this changed at recombination, when neutral atoms were formed and the universe became transparent to radiation. Photons interact with atoms much more weakly than they do with charged particles, so they are essentially free to propagate through the universe directly from the fireball, and eventually to our detectors⁴. We say that the photons decouple from matter. Thus when we look back as far as possible, to the epoch of recombination, we should see a panoramic “snapshot” of the universe as it was when its temperature was 3000 K. (This image of the infant universe is sometimes called the “surface of last scattering”, because the photons that make up the image arrive at our detectors after traveling on a straight path through space since the last time they were scattered during recombination). The peak wavelength of radiation at this temperature is near the red end of the visible spectrum, so a 3000 K fireball should glow with intense red light. Then why isn’t the sky red?

The reason is cosmological redshift. As photons propagate to us from the fireball, their wavelength is stretched by the expansion of the universe and is shifted far out of the visible range. At the same time, the density of photons is diluted by the expansion, so the radiation arrives at us highly red-shifted and with a strongly diminished intensity. If indeed the early universe was homogeneous and isotropic, the intensity of this relic radiation should be nearly the same in all directions on the sky.

11.4 Discovering the Primeval Fireball

Relic radiation from the primeval fireball was first predicted in the 1940’s by George Gamow’s two young colleagues, Ralph Alpher and Robert Herman. They estimated the present temperature of the radiation to be about 5 K. Detecting radiation of such a low temperature was a challenging task, and most observers at the time felt that it could not be done. So the prediction passed almost unnoticed (Fig. 11.6).

⁴This process is similar to how photons make their way from inside the Sun to the Earth. Photons that are inside the Sun (or any other star) are constantly scattered in random directions, and it can take millions of years for them to make their way to the surface of the Sun. Once they get there, they are no longer jostled about, and stream freely towards us, arriving within a mere 8 min.



Fig. 11.6 Ralph Alpher (right) and Robert Herman (left) predicted that relic radiation from a hot early epoch should pervade our universe. The word "Ylem" on the label of the bottle is the term invented by Gamow (depicted here as a genie coming out of a bottle) and his friends for the primeval fireball

More than a decade later, Robert Dicke at Princeton reinvented the idea of a hot primeval fireball and realized that it leads to the prediction of a pervasive cosmic radiation. Dicke assembled a group of three young physicists, assigning one of them, Jim Peebles, to work out the details of the theory and the other two, Peter Roll and David Wilkinson, to build a detector that would put the theory to test. Peebles was not aware of the work of Gamow's group, so he had to start from scratch. He completed the calculation in early 1965, predicting radiation with a thermal black body spectrum at a temperature of about 10 K. At that time the detector setup was also nearly complete, so the Princeton group was poised to either discover the primeval fireball or to prove that it never existed.

In the meantime, at Bell Telephone Laboratories in New Jersey, less than 50 km away from Princeton, Arno Penzias and Robert Wilson were testing a sensitive radio antenna that they hoped to use in a study of radio emission from the Milky Way. They first needed to account for possible sources of noise, such as radio emission from the Earth's atmosphere, and electronic noise in their antenna. But after half a year of work there still remained a persistent radio noise of unexplained origin. Penzias and Wilson measured

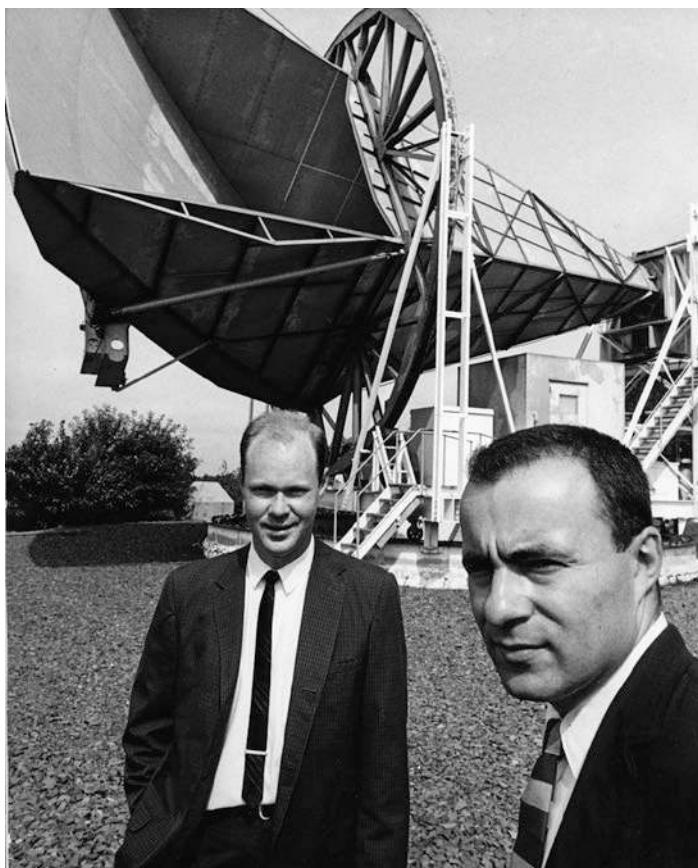


Fig. 11.7 Wilson (left) and Penzias (right) in front of their radio antenna. Credit AIP Emilio Segre Visual Archives, Physics Today Collection

the characteristic temperature of the noise to be about 3 K, corresponding to microwaves with a wavelength of 2 mm. The noise did not depend on the time of day or the direction in the sky, ruling out the atmosphere as a source. They also ruled out electronic noise and, believe it or not, pigeon droppings on the antenna! (Fig. 11.7).

Dicke and his group at Princeton were ready to start their measurements when they learned about Penzias and Wilson's predicament. They knew immediately that the mysterious noise whose origin puzzled Penzias and Wilson was precisely the signal of relic radiation that they were hoping to detect. The two teams published back-to-back papers in the same journal. Penzias and Wilson described their experiment, and the Princeton group interpreted it as a measurement of the cosmic radiation left over from the big bang.⁵ Today this radiation is called the cosmic microwave background (or CMB).

And what about Gamow's group? By the time of the CMB discovery none of them was actively working in cosmology. In the mid-1950s Gamow became interested in biology, where he contributed important insights into the genetic code, while Alpher and Herman moved on to careers in industry. Penzias and Wilson were awarded the Nobel Prize in 1978. No prize has ever been awarded for the prediction of the CMB.

Penzias and Wilson measured the intensity of the CMB at only one wavelength. To determine if this radiation was indeed part of a thermal spectrum, cosmologists still had to measure the radiation intensity over an extended range of wavelengths. This problem was tackled in a number of experiments in subsequent years, culminating in 1990 with the launch of NASA's Cosmic Background Explorer (COBE) satellite. COBE measured the spectrum of the CMB with unprecedented precision and found a perfect thermal spectrum (as predicted by the theory) with $T = 2.725\text{ K}$ (Fig. 11.8). Furthermore, the radiation intensity measured by COBE was nearly the same in all directions, with variations of less than 1/1000. Thus the early universe was indeed very isotropic and homogeneous.

11.5 Images of the Baby Universe

What do we actually see in the CMB? This depends on how accurately we measure the radiation temperature. If the accuracy is less than one part in 1000, then all we can see is a uniform radiation background, as in Fig. 11.9a. Here, the sky is represented with the so-called Mollweide projection, which is often used to represent the surface of the globe on a flat map.

⁵The observed value of the CMB temperature (3 K) is close to the theoretical prediction (5–10 K). The discrepancy between the two was mostly due to the uncertainty in the average matter density that was used in the calculations.

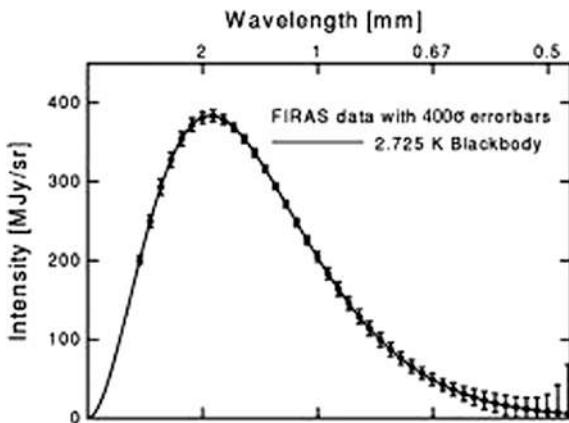


Fig. 11.8 COBE's measurement of the cosmic background radiation spectrum. The theoretical blackbody spectrum (solid curve) is superposed on the data points. The error bars have been magnified 400 times, so that they are visible

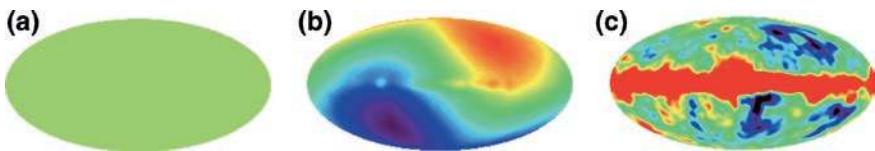


Fig. 11.9 COBE temperature pattern, at various levels of sensitivity. **a** Smooth background indicating the universe is homogeneous and isotropic. **b** Dipole pattern due to our motion relative to the CMB. In the direction we are heading, the CMB photons are slightly blueshifted (so their temperature is higher; hence this part of the sky is marked *red* in the figure, and in the opposite direction they are *redshifted* [so the corresponding part of the sky is *blue* (This choice of *color coding* can be confusing. Here the convention of everyday life, where *red* means “hot” and *blue* means “cold”, like on faucets, is followed. Unfortunately this convention is opposite to the fact that *blueshifted light* is hotter and bluer)]. **c** The *red band* comes from microwave emission in our galaxy. The other patches indicate tiny temperature fluctuations of the CMB, due to the presence of small density fluctuations at the time the CMB photons were emitted

At a somewhat higher accuracy, a “yin-yang” pattern emerges, as shown in Fig. 11.9b. Red coloring corresponds to higher and blue to lower than average temperature. This so-called dipole pattern is due to the motion of our Milky Way galaxy relative to the CMB. The highest temperature is observed in the direction of our motion, and the lowest temperature in the opposite direction. This is just the Doppler effect: as we move towards the incoming radiation, its wavelength shrinks and the temperature increases. The velocity of our motion through the CMB is about 600 km/s; it can be attributed to



Fig. 11.10 The COBE team leaders John Mather (left) and George Smoot (right) won the 2006 Nobel Prize in Physics for this work. Credit (for John Mather photo) NASA, courtesy AIP Emilio Segre Visual Archives, W. F. Meggers Gallery of Nobel Laureates. This image also available from NASA. Credit (for George Smoot photo) Photograph by Jerry Bauer, courtesy AIP Emilio Segre Visual Archives, W. F. Meggers Gallery of Nobel Laureates

the gravitational attraction of a large concentration of mass in the direction of the Virgo supercluster (Fig. 11.10).

The dipole component can easily be subtracted from the CMB temperature map. This reveals a pattern of higher and lower temperature regions shown in Fig. 11.9c. The red “equatorial” band comes from the microwave emission of our galaxy. In the rest of the map, the typical temperature variation between red and blue regions is only about one part in 100,000. These tiny variations reflect fluctuations in density. Higher density regions will later evolve into galaxies and galaxy clusters, as we will discuss in Chap. 12.

The temperature map in Fig. 11.9c was produced by the COBE satellite in 1992, after two years of taking data. It showed for the first time that the early universe did have small density fluctuations. However the resolution of COBE was rather limited, and much more work remained to be done to mine the vast amount of cosmological information contained in the CMB. Thus, NASA’s Wilkinson Microwave Anisotropy Probe (WMAP) satellite⁶ was launched in

⁶The satellite was named after David Wilkinson who played a major role in CMB research. (Remember, Wilkinson was one of the young fellows that Robert Dicke recruited to build a CMB detector in the 1960s).



Fig. 11.11 The COBE (launched 1989), WMAP (launched 2001), and Planck (launched 2009) satellites. Credit GSFC/NASA, NASA / WMAP Science Team, and ESA

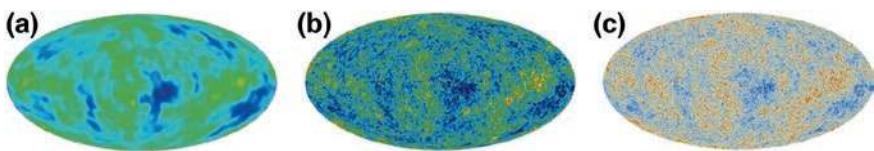


Fig. 11.12 CMB temperature maps of increasing resolution produced by a COBE, b WMAP, and c Planck satellites. Credit a COBE b NASA / WMAP Science Team c Copyright ESA and the Planck Collaboration

2001 to succeed COBE. While COBE's resolution was 7° (note the angular size of the full moon is 0.5°), WMAP had a resolution that is 33 times higher, and was 45 times more sensitive. WMAP operated successfully for 9 years, returning precision data that led to accurate determinations of the age, composition, and geometry of the universe. WMAP was succeeded by the Planck satellite, which had triple the resolution and was 10 times more sensitive. As the resolution improved, the image of the baby universe has become more and more focused (see Figs. 11.11 and 11.12). This image contains important information about physical phenomena that took place way before the epoch of recombination. We will have more to say about this later.

11.6 CMB Today and at Earlier Epochs

CMB photons are all around us, in huge numbers. Their number density (that is, the number of photons per cubic meter) is⁷ $n_\gamma \approx 4 \times 10^8 \text{ m}^{-3}$. This is comparable to the density of photons coming to us from a full moon.

⁷The Greek letter gamma, γ , is often used to denote photons.

If our eyes were sensitive to microwaves, we might be able to see in the cosmic light!

We can compare n_γ to the average number density of nucleons: $n_n \approx 0.24 \text{ m}^{-3}$. The nucleon to photon ratio is

$$\frac{n_n}{n_\gamma} \approx 6 \times 10^{-10} \quad (11.5)$$

which means that there are more than a billion photons present for every nucleon in the universe.

Even though microwave photons are much more numerous than nucleons, the energy (and equivalent mass) of each photon is very small compared to that of a nucleon. As a result, the mass density of the CMB is much smaller than the density of matter today

$$\frac{\rho_{\gamma 0}}{\rho_{m0}} \approx 1.7 \times 10^{-4}. \quad (11.6)$$

We now consider how the cosmic radiation evolves with time. As the universe expands, the wavelength of CMB photons grows in proportion to the scale factor, $\lambda \propto a$, and their energy decreases as $E \propto 1/a$. Since the wavelengths of all photons are stretched by the same factor $a(t)$, the thermal form of the radiation spectrum is preserved. The radiation temperature T is proportional to the average energy of photons; hence

$$T \propto 1/a. \quad (11.7)$$

It follows from Eq. (11.7) that

$$\frac{T}{T_0} = \frac{1}{a}. \quad (11.8)$$

Here, $T_0 \approx 3 \text{ K}$ is the present CMB temperature and we use the convention that the scale factor is $a_0 = 1$ at the present time. This useful formula allows us to determine how much the universe has expanded since the time when it had temperature T . For example, the temperature at recombination is $T_{rec} \approx 3000 \text{ K}$, and we find from Eq. (11.8) that the scale factor at that time was $a_{rec} \approx 10^{-3}$. This means that the universe has expanded by a factor of 1000 since the time of recombination. The corresponding redshift is $z_{rec} \approx 1000$.

The cosmic time t_{rec} at recombination can be found by solving the Friedmann equation for the scale factor $a(t)$. This gives $t_{rec} = 380,000$ years.

The CMB thus provides an image of the universe at 380,000 years after its birth—a very early time, compared to the present cosmic age of about 14 billion years.

In the course of cosmic expansion, the number density of photons is diluted as

$$n_\gamma \propto \frac{1}{V} \propto a^{-3}, \quad (11.9)$$

where $V \propto a^3$ is the volume of an expanding region. At the same time, the energy of each constituent photon decreases as $E \propto a^{-1}$, due to redshift. The overall effect is that the radiation energy density is proportional to

$$\rho_\gamma \propto a^{-4}. \quad (11.10)$$

11.7 The Three Cosmic Eras

The energy density of the universe is now dominated by dark energy (69%), it has a substantial matter component (dark 26% and atomic 5%), and trace amounts of radiation. However, the three energy components evolve in different ways: the matter and radiation densities decrease respectively as $\rho_m \propto a^{-3}$ and $\rho_\gamma \propto a^{-4}$, while the vacuum energy density remains constant. As a result, the composition of the universe at earlier times was rather different from what it is now (see Fig. 11.13).

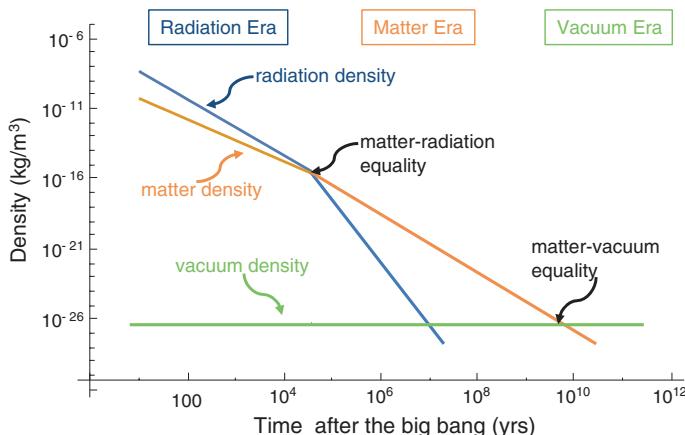


Fig. 11.13 The evolution of energy density for each cosmic component

Even though ρ_γ is much smaller than ρ_m today, as we follow cosmic evolution back in time, the radiation density grows faster than that of matter, and at $t_{eq} \approx 60,000$ years the two densities become equal (see the Appendix). At earlier times the radiation density was larger than both the matter and dark energy density. Thus this era is called the radiation era. At later times, the matter density dominates over radiation and dark energy. This matter era lasts for several billion years. Finally, roughly 5 billion years ago, the matter density dropped below that of the vacuum energy density, and the universe entered its current vacuum dominated era.

During the radiation and matter era, the scale factor grows as $a(t) \propto \sqrt{t}$ and $a(t) \propto t^{2/3}$, respectively (as we show in the Appendix). In both of these eras the horizon distance grows linearly with time, $d_{hor} \sim t$, which is faster than $a(t)$. Thus more and more of the universe becomes visible with time. The situation, however, is different in the vacuum era where the scale factor grows faster than the horizon. In this case, more and more of the observable universe exits our horizon, and less and less of the universe becomes visible with time (we will return to this in Chap. 16).

Summary

When we follow the expansion of the universe backwards in time, the density and temperature increase without bound. All structure disintegrates at high temperatures; thus the early universe starts out as a fireball of the most basic particles, including electrons, protons, neutrons, and photons. The fireball uniformly fills the whole universe. As the universe expands, it cools down, and composite objects begin to form. Within the first three minutes after the big bang, the temperature dropped sufficiently for protons and neutrons to bind together into atomic nuclei. Then at about 380,000 years, electrons and nuclei combined to form neutral atoms, and the universe became transparent to light. We can now detect the radiation emitted from the fireball at that epoch; it comes to us from all directions in the sky. This is what we call the cosmic microwave background radiation.

In broad-brush terms, the history of the universe can be divided into three cosmic eras: the radiation era, the matter era, and the current dark energy era. During each era the energy density is dominated by radiation, matter and dark energy, respectively.

Questions

- As we follow the universe backwards in time, what happens to its temperature T and density ρ as we approach the big bang?

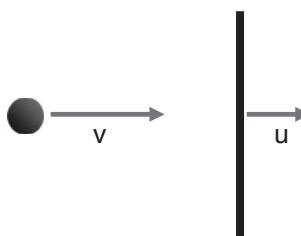


Fig. 11.14 Particle and retracting wall

2. What happens to the temperature of gases as they expand? And as they are compressed?
3. Consider a particle moving at speed v towards a wall which is retracting at speed $u < v$ in the same direction, as shown in Fig. 11.14. What will be the particle's speed after it bounces off the wall? (Hint: consider how this process looks to an observer moving with the wall, who sees the particle bounce back at the same speed as it arrives.)
4. Today the universe is filled with complex structures like atoms and molecules. What happens to these structures as we go far back to the hot early universe?
5. Is it possible to break nuclei into separate protons and neutrons? Would this happen at a higher or lower temperature than the ionization of hydrogen? (An atom of hydrogen is “ionized” when its electron is separated from the nucleus.)
6. If photon A has twice the wavelength of photon B, by how much is its energy greater or less than that of photon B?
7. Briefly explain the relation between the rate of photon emission and absorption for a system that is in thermal equilibrium.
8. The Sun is a thermal emitter with a surface temperature of 6000 K. If the temperature of the Sun’s surface were to double, what would happen to the intensity of the sun’s radiation?
9. The peak intensity of radiation of the Sun is right in the middle of the visible part of the spectrum. Do you think this is just a coincidence?
10. Did the initial fireball explode into empty space? Please explain.
11. What is recombination? What is the “surface of last scattering”?
12. As we observe distant objects in the universe, do we see them as they are today, or as they were at some time in the past? Why? Does this help or hinder us in our quest to understand the evolution of the universe?

13. Why can we not look all the way back to the big bang?
14. The microwave background radiation was emitted when the temperature of the universe was 3000 K. Objects in thermal equilibrium at this temperature glow red, so why are we surrounded by a sea of microwave photons, instead of red photons?
15. What happens to the wavelength of photons as they propagate in an expanding universe?
16. Is there any way to account for the CMB radiation in the steady state theory?
17. Does the CMB spectrum shape (the solid curve in Fig. 11.8) vary from one direction in the sky to another? Does it change as the universe evolves? Why?
18. Explain in what sense the CMB provides an image of the universe at 380,000 years after the big bang.
19. What do the differences in color on the CMB maps in Fig. 11.12 represent?
20. If someone claimed to discover a galaxy at a redshift of 2000, would you believe it? Why/Why not?
21. Why is the CMB slightly hotter in one half of the sky than in the other?
22. Consider a galaxy at a redshift of $z = 1$. What was the average matter density at the time light left the galaxy compared to today? What was the average energy density of radiation then compared to today? What was the temperature of the CMB then?
23. By how much has the universe grown in size since the time when matter and radiation densities were equal to each other? (Hint: you may use the present value of the ratio $\frac{\rho_r}{\rho_m}$ in Eq. (11.6) and figure out how this ratio depends on the scale factor a .) What was the temperature of the universe at the time of equal matter and radiation densities?
24. If there are any observers around in half a trillion years from now, will they be able to see other galaxies? Will they see CMB radiation? If so, how would that radiation be different from what we observe today?

12

Structure Formation

The universe is lit up with stars, which are scattered through space, forming a hierarchy of structure. Stars are assembled into galaxies, and galaxies are grouped in clusters, which are in turn grouped into still larger structures, called superclusters. The formation of cosmic structures is an active area of research, and we now believe we have a good idea of how they emerged.

12.1 Cosmic Structure

Galaxies come in three main types: spiral, elliptical and irregular, as shown in Fig. 12.1. The main components of a spiral galaxy are the central bulge, the flattened disk with spiral arms, and a huge dark matter halo. The disk of our Milky Way galaxy is roughly 100,000 light years across and about 10,000 light years thick. The halo is nearly spherical, with a diameter about ten times larger than that of the disk. The Sun sits in the disk and is located about 25,000 light years from the galactic center. Large galaxies like the Milky Way contain of order 100 billion stars. The typical distance between stars in a galaxy is a few light years; this is much larger than the size of a star. If you imagine the Sun to be pea sized, the nearest star would be 160 km away! Thus galaxies are mostly empty.

Galaxies group together to form clusters. The Milky Way belongs to a small cluster called the Local Group (see Fig. 12.2). The Andromeda galaxy also resides in the Local Group, some 2.5 million light years away. Although



Fig. 12.1 Spiral, elliptical and irregular galaxies. Spiral Image (NGC 6814) Credit ESA/Hubble & NASA; Acknowledgement Judy Schmidt (Geckzilla). Elliptical (M87) Credit Canada-France-Hawaii Telescope, J.-C. Cuillandre (CFHT), Coelum. Irregular (NGC 1427A) Credit NASA, ESA, and The Hubble Heritage Team (STScI/AURA)

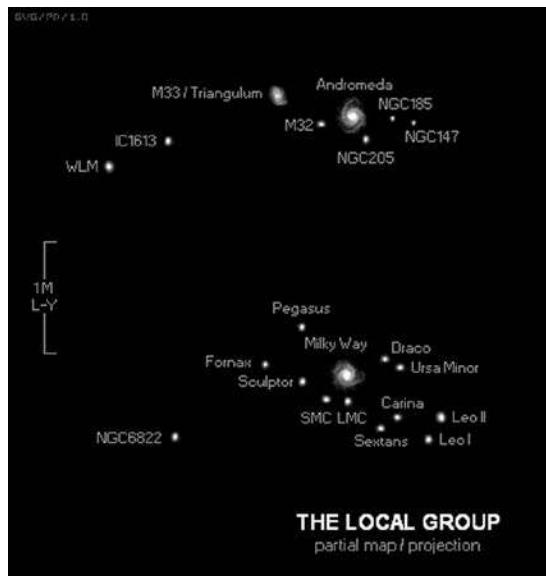


Fig. 12.2 The Local Group. Credit www.wikipedia.org

the Local Group has less than 40 galaxies, some rich clusters contain thousands of galaxies. The closest cluster to the Local Group is the Virgo cluster, which lies about 60 million light years away and contains over a thousand galaxies (Fig. 12.3).

Clusters are further grouped into superclusters, some of which contain hundreds of clusters (the Local Group is part of the Local Supercluster).

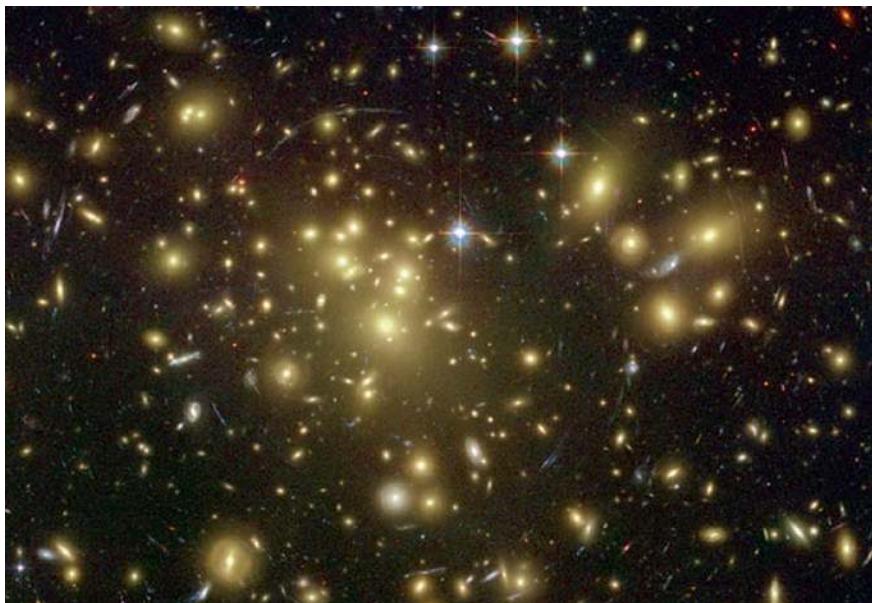


Fig. 12.3 The virgo cluster of galaxies (by the Hubble Space Telescope). Credit NASA/ESA

Automated galaxy surveys, pioneered in the late 1980s at the Harvard Center for Astrophysics (CfA), also revealed that the galaxy distribution has a frothy appearance, with filaments and sheet-like walls of galaxies straddling huge voids (Fig. 12.4).

What happens on still larger scales? Does the hierarchy of structure continue to grow, with superclusters grouping together and so on? Or does the distribution of galaxies become uniform above a certain scale? The Anglo-Australian Observatory “Two Degree Field” (or 2dF) galaxy survey and the Sloan Digital Sky Survey (SDSS) set out to answer these questions. Extending out to 2 billion light years, these enormous surveys show that the large-scale distribution of galaxies exhibits a web-like structure with filaments, sheets, clusters and voids (see Fig. 12.5). Consistent with the CfA results, the largest structures are roughly 300 million light years in size, and have a mass of about 10^{17} Solar masses (or 10^5 galactic masses). On still larger scales, the universe is homogeneous. There are no “super-superclusters”. Thus, if the matter distribution were smoothed over distance scales of 300 million light years, the universe would be homogeneous and isotropic, as assumed in the Friedmann models.

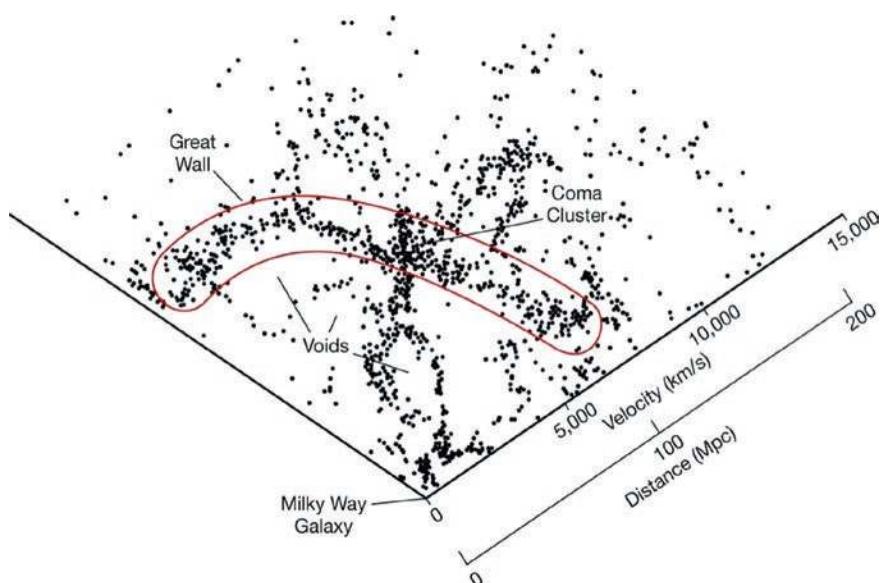


Fig. 12.4 CfA map of a thin (6°) slice of the universe. Each dot represents a galaxy. Some of the apparently filamentary structures in the map are actually slices through sheet-like walls. One of them is the "Great Wall" outlined in red on the map. Credit Smithsonian Astrophysical Observatory. De Lapparent, V.; Geller, M. J.; Huchra, J. P. (1986). "A slice of the universe". The Astrophysical Journal. 302: L1

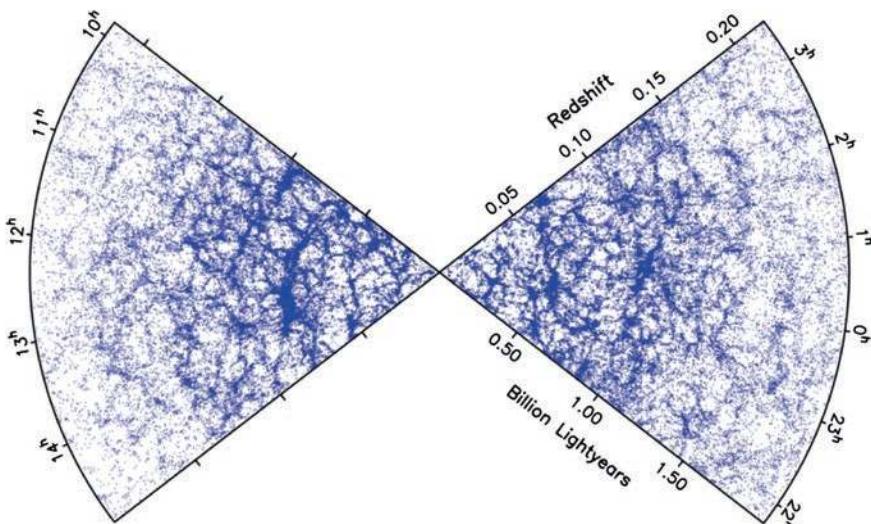


Fig. 12.5 The large-scale distribution of galaxies, as observed by the 2dF survey (2002). Credit 2dF Galaxy Redshift Survey

12.2 Assembling Structure

Now that we are familiar with the hierarchical structure of the universe, we are faced with an inevitable question: *How did all these structures arise?* We know that the universe was very homogeneous when the CMB radiation was produced, at $t_{rec} = 380,000$ years. If it were perfectly homogeneous, having the same density everywhere, then it would remain that way forever. But we know from CMB observations that there were tiny deviations from perfect homogeneity (about one part in a hundred thousand). This is all that was needed to seed structure formation.

A region that is denser than average will attract matter gravitationally from its surroundings. It thus gets even denser relative to the average, and its gravitational pull gets stronger, attracting even more matter. An over-dense expanding region initially continues to expand, but eventually it turns around and collapses back on itself, forming a gravitationally bound object. This effect is called *gravitational instability*; it causes the matter distribution to become more and more lumpy. Only a tiny fluctuation is needed to get the process started.

The basic idea is simple, but as it often happens, the details are rather complicated. It took several decades to sort them out, ultimately the following picture has emerged:

- The gravitational instability is effective only if the expansion of the universe is sufficiently slow. During the radiation era the universe is expanding too fast, so the growth of density fluctuations can begin in earnest only at $t_{eq} \approx 60,000$ years, when the matter era begins.
- The clustering of atomic matter develops differently from that of dark matter. The hot atomic gas has a high pressure, which prevents it from being pulled into clumps less massive than about 10^6 Solar masses. On larger mass scales, gravity wins and pressure does not play an important role. Dark matter, on the other hand, is influenced only by gravity,¹ so its lumpiness begins to grow right after t_{eq} . (Here we assume that the dark matter is “cold”, in the sense that its particles move slowly and cannot escape the gravitational pull of the developing clumps. This is naturally satisfied if the dark matter particles are sufficiently massive).

¹Another difference of atomic gas from dark matter is that gas particles often collide, emitting photons in the process. As a result the gas loses energy, cools and sinks deeper towards the centers of dark matter clumps. This cooling process is important on galactic mass scales, up to 10^{12} Solar masses. Dark matter particles, on the other hand, interact very weakly and lose almost no energy in collisions. This explains why stars and gas are localized near the centers of dark matter halos.

- Smaller clumps take a shorter time to form; as a result structure formation proceeds in a hierarchical, bottom-up fashion. It begins with the formation of small dark matter clumps, which then merge to form larger and larger structures. Atomic matter starts falling into dark matter clumps at about 0.5 billion years ABB² (after the big bang), when the typical mass of a clump exceeds 10^6 Solar masses. This is when the first stars are formed. Their light illuminates the universe, ending the cosmic dark age.
- A spherical overdense region would collapse to a localized gravitationally bound object. But a typical overdense region is more like an ellipsoid, which has different sizes along three orthogonal axes. It first collapses along its smallest axis to form an approximately 2-dimensional sheet. The sheet then collapses to form a filament, and finally the filament collapses to a localized halo. The observed galaxies and clusters are well localized, gravitationally bound objects, while superclusters are caught in the process of their formation. Some of them resemble filaments or sheets, while others have a rather irregular appearance.
- About 5 billion years ago, when the matter era gave way to the current vacuum dominated era, the expansion of the universe started to accelerate, causing further gravitational clumping to be quenched. This is why there will never be cosmic structures that are larger than superclusters.

It is interesting to note that dark matter played a crucial role in structure formation. In the absence of dark matter, density fluctuations would start growing only much later than t_{eq} . The amount of growth that could occur by the onset of the vacuum era would then be insufficient for the formation of bound structures. Thus, if it were not for dark matter, the universe would be almost devoid of galaxies today.

12.3 Watching Cosmic Structures Evolve

The hierarchical scenario of structure formation has been tested by direct observation of distant galaxies. We can see what early galaxies looked like by taking galaxy images at higher and higher redshifts, that is, by looking deeper and deeper into space. The image in Fig. 12.6 was obtained by pointing the Hubble Space Telescope to a totally blank spot in the sky and tak-

²We will use the notation ABB to mean “after the big bang” throughout the rest of the book.

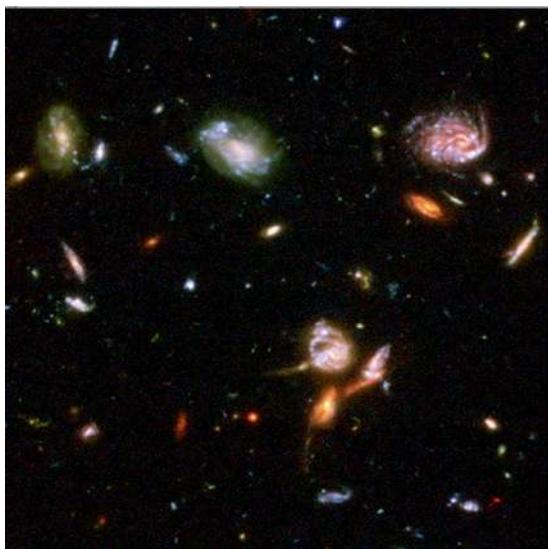


Fig. 12.6 The Hubble Ultra Deep Field *Credit* NASA; ESA; G. Illingworth, D. Magee, and P. Oesch, University of California, Santa Cruz; R. Bouwens, Leiden University; and the HUDF09 Team

ing a very long exposure. This is known as the Hubble Ultra Deep Field. It shows a multitude of early galaxies—some of them date back to less than a billion years ABB.

These infant galaxies differ from today's galaxies in a number of ways. They are much smaller, with a typical size of 10,000 light years across, and generally appear to be chaotic and irregular. Furthermore, almost all infant galaxies are colliding or interacting gravitationally with their neighbors (see Fig. 12.7), compared to about 2% that are colliding today. All these features strongly suggest that present-day galaxies formed by collisions and mergers of smaller early galaxies.

Another way to see cosmic structure formation unfold in front of you is to use a computer simulation. A few frames from one such simulation are shown in Fig. 12.8. The simulation follows the history of a large cubic volume as it expands to the present size of 160 million light years. It starts with a nearly uniform distribution of particles in the volume and includes only gravitational interactions of the particles and the gravitational effect of dark energy. This is a good approximation on the largest scales (galaxy clusters



Fig. 12.7 Colliding early galaxies *Credit* NASA, ESA, the Hubble Heritage (STScI/AURA)-ESA/Hubble Collaboration, and A. Evans (University of Virginia, Charlottesville/NRAO/Stony Brook University)

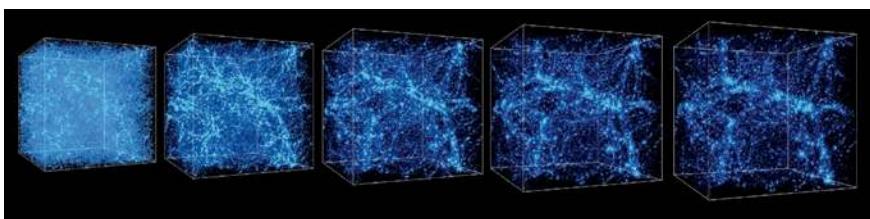


Fig. 12.8 Computer simulation of structure formation. The evolution of a cubic volume is followed from 0.5 billion years ABB ($z = 10$) to the present epoch. The cube expands with the universe (by a factor of 11), but the expansion has been factored out in the figure, so all “snapshots” of the cube have the same apparent size. The snapshots are at redshifts 10, 4, 1, 0.5 and 0. (This simulation was performed by A. Kravtsov and A. Klypin at the National Center for Supercomputer Applications at the University of Chicago)

and above), where gravity is the dominant force. The web-like distribution of matter in the last frame closely resembles the large-scale galaxy distribution observed today.

On galactic and smaller scales, the complicated dynamics of atomic gas cannot be ignored. Cosmologists are making progress simulating these dynamics on computers, but some details of galaxy and star formation are still not fully understood. This is now an active area of research.

12.4 Primordial Density Fluctuations

The picture of structure formation by gravitational instability relies on the existence of small primordial density fluctuations. The magnitude of these fluctuations can be different for regions of different size (or mass). In order

to fully characterize the fluctuations, one has to specify their *spectrum*—that is, the typical strength of the fluctuation as a function of mass. If the spectrum is peaked at some particular mass, then the first bound objects to form are likely to have this mass.³ Thus the evolution of structure in the universe depends on the form of the primordial fluctuation spectrum.

The structure formation scenario outlined in this chapter and computer simulations, like the one in Fig. 12.8, assume the fluctuation strength to be approximately the same for all relevant mass scales. This is called the *scale-invariant* fluctuation spectrum. *But what is the origin of the primordial fluctuations? And what determined their spectrum?* These questions are addressed by the theory of cosmic inflation. We shall see in Chap. 17 that the fluctuation spectrum predicted by this theory is indeed nearly scale-invariant. Moreover, we shall see that this form of the spectrum is supported by observations.

12.5 Supermassive Black Holes and Active Galaxies

General relativity predicts that if we cram a large enough amount of mass into a small enough volume, we can create a black hole (see Chap. 4). Stellar mass black holes can form at the final stages of stellar evolution. There is also strong observational evidence for the existence of colossal, supermassive black holes with masses of millions and even billions of Solar masses. Velocities of stars and gas close to galactic centers have been measured using Doppler shifts. These measurements reveal the presence of extremely massive compact objects lurking at the center of most galaxies—they are black holes! The black hole at the center of our Milky Way has a mass of about 3.7 million Solar masses, while black holes in some other galaxies are more massive than a billion Solar masses.

These monstrous black holes lie dormant most of the time. But when there is some gas near the galactic center, it falls into the black hole. The gas heats up and emits vast amounts of radiation as it spirals down into the hole. Radiation continues until the supply of gas is exhausted. During their explosive periods, supermassive black holes are the most luminous objects in the universe. Depending on the amount of energy released, and the type of radiation, we call them quasars or active galactic nuclei (Fig. 12.9).

³There are also some additional factors that determine the mass (or size) of the first collapsed objects; we do not need to discuss this here in more detail.

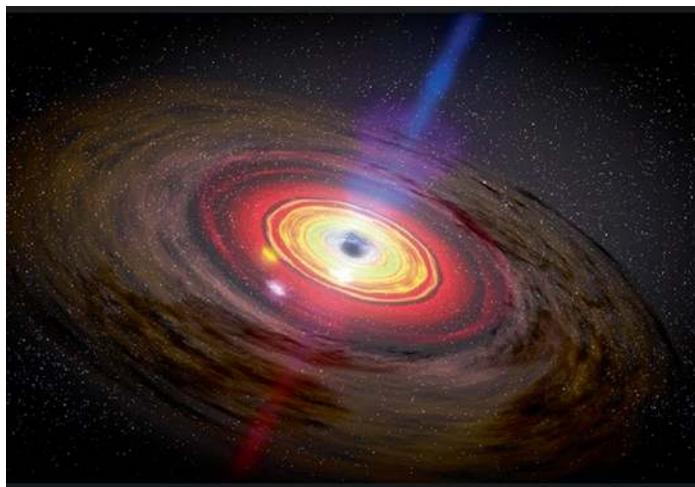


Fig. 12.9 Artist's impression of a supermassive black hole. Infalling gas forms a thin accretion disk as it spirals into the hole. Powerful jets of gas emanate from the black hole vicinity in the directions orthogonal to the disk. Such jets are often observed in active galactic nuclei, but the mechanism of their formation is not well understood. *Credit NASA / Dana Berry, SkyWorks Digital*

Galaxy collisions perturb the gas distribution in the galaxies and trigger black hole feeding. Thus black hole activity should have been more common in the young universe, when the collision rate was high. Observations of quasars support this hypothesis. The quasars we observe are very distant, thus they are very old. Quasar formation rates peaked at about 2–3 billion years ABB. They represent an early stage in galaxy development.

Summary

The origin of galaxies can be traced to tiny inhomogeneities in the primordial fireball. Some regions of the fireball had a slightly higher density than others. The mass of these regions grew as they attracted matter from the surrounding space, and over the course of billions of years they evolved into galaxies and larger structures.

The scenario of structure formation suggested by combining mathematical analysis with simulations is that the first stars formed around 0.5 billion years after the big bang, followed by the birth of infant irregular galaxies. Large galaxies were formed via hierarchical assembly, from the merger of smaller ones. Galaxies then clumped into clusters, which further grouped into superclusters. On the largest scales, the universe exhibits a web-like

structure with filaments, sheets, and voids, up to scales of roughly 300 million light years. On still larger scales, the growth of structure is quenched by gravitational repulsion due to dark energy, so there is no further clumping, and the universe is homogeneous.

Questions

1. The largest structures in the universe have sizes of about 300 million light years. On still larger scales what does the universe look like? What does the galaxy distribution look like on smaller scales?
2. What is gravitational instability? Describe how it can turn small density fluctuations into structures like stars and galaxies.
3. Describe how gravitational instability is different for dark matter and for atomic gas.
4. What is hierarchical clustering?
5. Can you explain why the large-scale distribution of galaxies has a web-like appearance, with filaments, walls and voids?
6. What observational evidence do we have for hierarchical structure formation?
7. Given that the radius of a Solar mass black hole is 3 km, what is the radius of a billion Solar mass supermassive black hole? Compare it to the radius of Earth's orbit around the Sun (1.5×10^{11} m). (Recall the formula for the Schwarzschild radius in Chap. 4).
8. Imagine you live at an earlier cosmic epoch, when galaxies were much younger than they are today. How would the galaxies that you observe be different from the present galaxies? Would they be larger, smaller, or about the same size? Would they be closer or farther apart? What other differences would you expect?
9. What is a quasar? Why are most of the quasars observed at early cosmic times?
10. How do we deduce that there is a supermassive black hole in the center of a galaxy? How do we measure the mass of black holes?
11. How do cosmologists map out the 3-dimensional distribution of galaxies from the 2-dimensional pattern observed on the sky?

13

Element Abundances

The chemical composition of the universe is rather simple. About 75% (by mass) of atomic matter is in the form of hydrogen, and almost all the rest is in the form of helium. All other chemical elements contribute less than 2% of the atomic mass. *Why is it that some elements are more abundant than others? And where did the elements come from in the first place?* These questions cannot be avoided, since atoms, and even atomic nuclei, could not exist in the early moments after the big bang. The challenge, then, is to understand how the elements were created by physical processes during the course of cosmic evolution.

13.1 Why Alchemists Did Not Succeed

The chemical properties of an atom are determined by the number of electrons it contains, which is equal to the number of protons residing in the nucleus. The number of protons defines the type of chemical element. For example, hydrogen has 1 proton, and gold has 79 protons. Atoms that have the same number of protons but different numbers of neutrons have almost identical chemical properties; they are called isotopes. For example, helium-4 has two protons and two neutrons in the nucleus, while helium-3 has two protons and only one neutron in the nucleus. The composition of some of the simplest atomic nuclei is shown in Fig. 13.1. There are 92 naturally occurring elements; their relative abundances in the universe are plotted in Fig. 13.2.

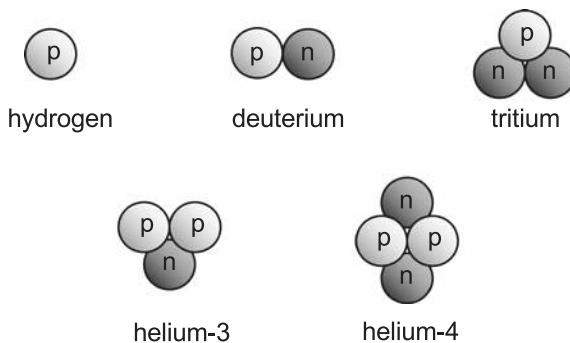


Fig. 13.1 The simplest atomic nuclei, with protons and neutrons represented by p and n, respectively. Deuterium and tritium are isotopes of hydrogen, while helium-3 and helium-4 are isotopes of the chemical element helium

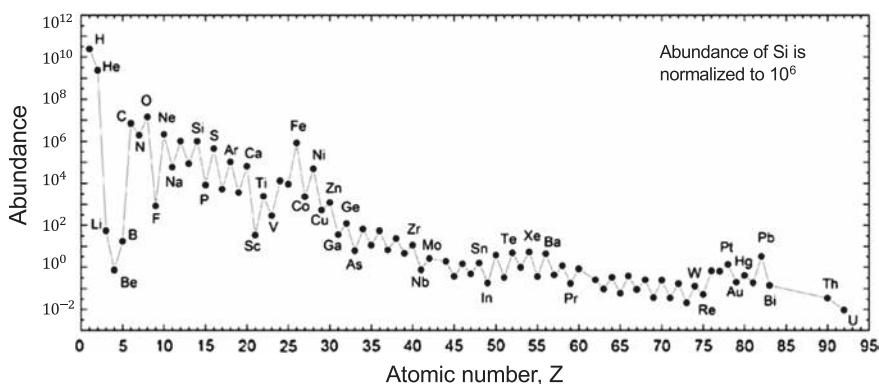


Fig. 13.2 Element abundances. The atomic number Z equals the number of protons in the nucleus

Alchemists in the Middle Ages tried to turn more abundant elements into gold. Newton also devoted much of his time to research in alchemy. Today we know there was a good reason why this research was doomed. In order to change one chemical element into another, one has to learn how to change the number of protons in the atomic nuclei. There are at least two ways to do so. First, you can hit the nucleus with something, so that it splits into two; and second, you can smash two nuclei together, hoping that they will merge and form a larger nucleus. Both methods have their problems.

The problem with the first method is that protons and neutrons are held in the nucleus by the strong nuclear force, so it takes a collision of very high energy to break a nucleus into two. The problem with the second method

is that the attractive nuclear force is strong only at very short distances, so before you can make two colliding nuclei stick together, you have to bring them very close to one another. There is, however, an electric repulsion between positively charged nuclei, and once again you need to supply very high energy to overcome this repulsion. The particle energies needed for nuclear transformations require temperatures in excess of tens of millions of degrees Kelvin. Such temperatures are naturally reached only in stellar interiors and in the early universe.

George Gamow, the founder of the hot big bang theory, suggested that elements were synthesized shortly after the big bang. He developed this idea in collaboration with Ralph Alpher and Robert Herman. It is now called *big bang nucleosynthesis*.

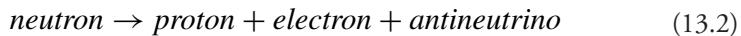
13.2 Big Bang Nucleosynthesis

Let us go back to a time around the first few seconds after the big bang. The temperature of the universe is several billion Kelvin and the fireball is a mix of neutrons, protons, electrons, photons, and neutrinos. At earlier times, neutrons and protons were in thermal equilibrium, converting back and forth into one another through weak nuclear processes like



A neutron is more massive than a proton and electron combined, so it costs extra energy to convert a proton into a neutron. As the universe expands and cools, energetic electrons needed to make the conversion become more and more rare, and neutrons become less and less numerous relative to the protons. At about one second ABB, the rates of conversion reactions become too slow to keep up with the expansion of the universe, so proton to neutron conversions effectively cease. At that time, there are approximately 6 protons for each neutron.

Isolated neutrons are unstable¹ and decay into protons and lighter particles with an average lifetime of 15 min:



¹Neutrons in atomic nuclei are stabilized by strong nuclear interactions.

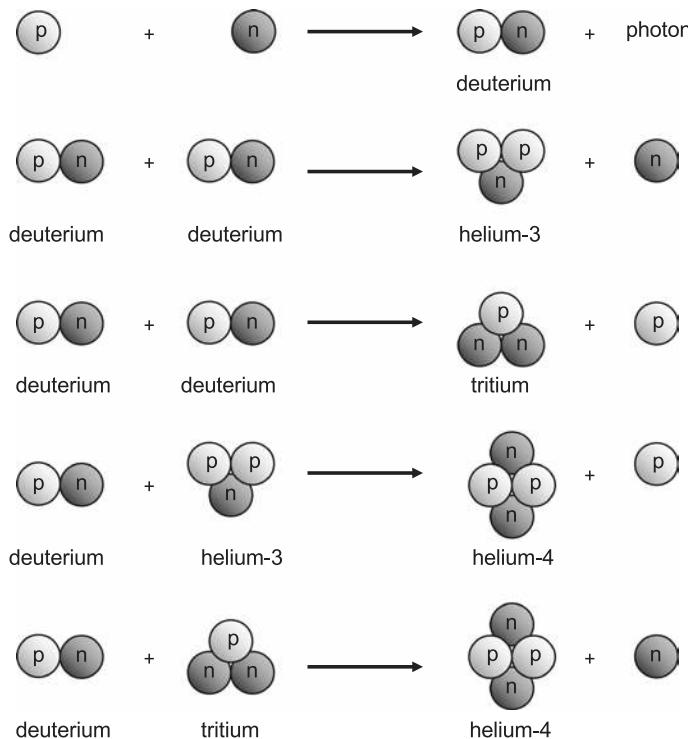


Fig. 13.3 The main nuclear reactions in big bang nucleosynthesis

Thus the universe had about 15 min to create the elements. Otherwise, it would have run out of neutrons, and the only element it would contain would be hydrogen.

The first step in nucleosynthesis is for a neutron to fuse with a proton to make deuterium, or heavy hydrogen (see Fig. 13.3). The deuterium is very fragile, but it is possible for two deuterium nuclei to fuse to make either helium-3 plus a neutron, or tritium plus a proton. There are now two ways in which helium-4 (this is the common helium nucleus consisting of two neutrons and two protons) can be formed: deuterium can fuse with helium-3 to make a stable helium-4 nucleus (plus a proton), or deuterium can fuse with tritium to form a stable helium nucleus (plus a neutron).

Early on, the main impediment to this chain of reactions is that the fragile deuterium nucleus gets destroyed by collisions with energetic photons before it combines with more deuterium to form helium-3 or tritium. But at about 1 min ABB, when the temperature dropped to one billion Kelvin, the photon energies are no longer sufficient to break deuterium. By this

time, the neutron to proton ratio has dropped to 1:7. From this point on, nucleosynthesis proceeds rather quickly, until almost all neutrons end up in helium. The process is essentially complete when the universe is about 3 min old. The abundance of helium predicted in this scenario is 25% by mass,² in excellent agreement with observation. The bulk of the remaining atomic mass is in the form of hydrogen.

Trace amounts of other light elements were also created during big bang nucleosynthesis. The predicted abundances of these elements are sensitively dependent on the nucleon density n_n at that epoch. Since the density changes with the expansion of the universe, it is more convenient to use the nucleon to photon ratio, $\eta = \frac{n_n}{n_\gamma}$, which is nearly independent of time. Good agreement between theoretical predictions and the observed abundances (about 10^{-5} deuterium, 10^{-5} helium-3, and 10^{-10} lithium-7) is achieved for $\eta \approx 6 \times 10^{-10}$ (see Fig. 13.4)³. Thus we should expect the universe to have about 1.6 billion photons for each nucleon.

This result has far reaching implications. Using the value of η and the present number density of CMB photons ($n_\gamma \approx 4 \times 10^8 \text{ m}^{-3}$, see Sect. 11.6), we can find the present average number density of nucleons, $n_n = \eta n_\gamma \approx 0.24 \text{ m}^{-3}$. Furthermore, considering that almost all the mass of atomic matter comes from nucleons (they are about two thousand times heavier than electrons), we can determine the average atomic mass density, $\rho_{at} \approx n_n m_n \approx 4 \times 10^{-28} \text{ kg/m}^3$, where $m_n \approx 1.7 \times 10^{-27} \text{ kg}$ is the nucleon mass. In terms of the critical density, this gives $\Omega_{at} = \rho_{at}/\rho_c \approx 0.05$. This value is comparable to the observed amount of matter in stars and interstellar gas—but then not much is left to account for the dark matter.

The amount of dark matter in the universe is about five times that in stars and gas, $\Omega_{dark} \approx 0.26$. Now we have to conclude that most of this matter cannot be made of ordinary atoms. It must consist of some “exotic” stable particles that have not yet been discovered.

The agreement between theory and observation for the abundances of the light elements is a remarkable accomplishment. But what about the heavy

²This number is easy to understand, considering that (i) there are 7 protons for each neutron (or 14 protons for 2 neutrons) and (ii) almost all neutrons end up in helium. For each helium nucleus (capturing 2 neutrons and two protons), 12 protons (hydrogen nuclei) are left free. Therefore, by mass, there is a 12:4 ratio of hydrogen to helium, which is precisely the 75% hydrogen, 25% helium prediction.

³The lithium abundance obtained by more accurate recent measurements is lower than predicted by about a factor of 3. The origin of this discrepancy is presently unclear; it is a subject of intense investigation.

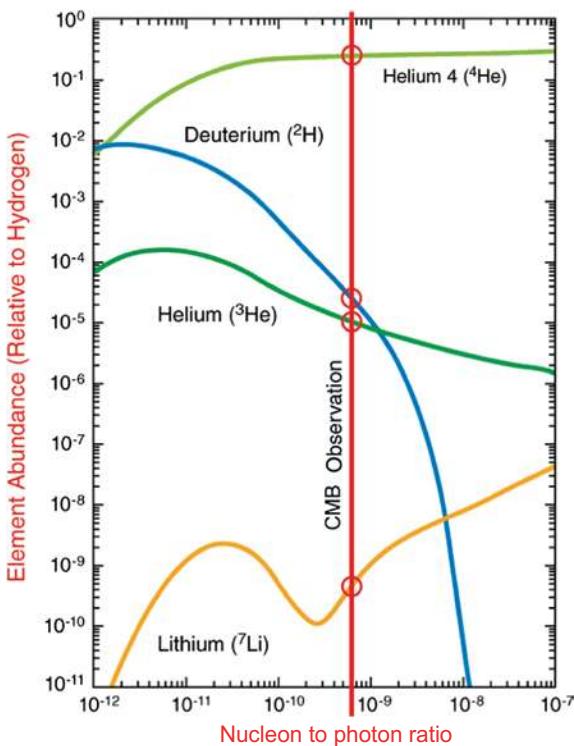


Fig. 13.4 Light element abundances. Solid curves indicate theoretical predictions depending on the nucleon to photon ratio. Observed element abundances are indicated by circles. There is only one value for the nucleon to photon ratio (red vertical line), which passes through all the data points. This value agrees with an independent measurement based on the CMB observations. Credit NASA/WMAP Science Team

elements? Here Gamow, Alpher and Hermann reached an impasse. The big bang nucleosynthesis does not progress much beyond helium. The reasons are that there are no stable nuclei consisting of 5 nucleons, and that a simultaneous attachment of two or more nucleons is highly unlikely. This “shortcoming” of the big bang model, known as the 5-nucleon gap, allowed other cosmological theories to thrive as viable alternatives, for a while.

In particular, before the big bang model was generally accepted, Fred Hoyle, Hermann Bondi and Thomas Gold proposed the *steady state* theory (as discussed in Chap. 7). According to this theory, the state of the universe has not changed over time, and the universe never went through a hot explosive stage. Hoyle suggested therefore that all elements were created in stars. While we now know that the hydrogen and helium abundances are set during the big bang, Hoyle was partly right about nucleosynthesis: elements heavier than lithium were indeed formed in stars.

13.3 Stellar Nucleosynthesis

Stars are gaseous spheres held together by gravity and heated by nuclear reactions in their interiors. Our Sun is a typical middle-size star, consisting mostly of hydrogen (71%). Its surface temperature is 6000 K and the central temperature is 10^7 K. Hydrogen is burned into helium in the central parts of stars like the Sun; helium ash is collected at the core. When all the hydrogen is burnt in the central region, the star can no longer support itself against gravity. The core begins to contract, and its temperature rises. Outside the core, a shell of hydrogen continues to burn into helium. Once the core reaches a temperature of $T \sim 10^8$ K, the helium ash starts burning to carbon and oxygen.

For a star with about the Sun's mass, nuclear reactions do not go beyond this point. Due to the heat generated in the core, the star will swell to become a red giant, and then will blow off its envelope, leaving a compact white dwarf remnant. If such a remnant happens to be in a binary pair, then it is possible for the white dwarf to accrete matter from its partner. Once a critical amount of matter is added to the star, it can undergo a supernova explosion. On the other hand, for massive stars (8 or more Solar masses), core burning can continue all the way until iron is formed. The process stops at iron, which is the most stable of all nuclei (see Fig. 13.5). Thus, more massive stars have a layered structure, with heavier elements produced at deeper levels, where the temperature is higher.

When a massive star runs out of nuclear fuel at the center, the central core collapses, reaching enormous densities and temperatures $T \sim 10^{10}$ K.

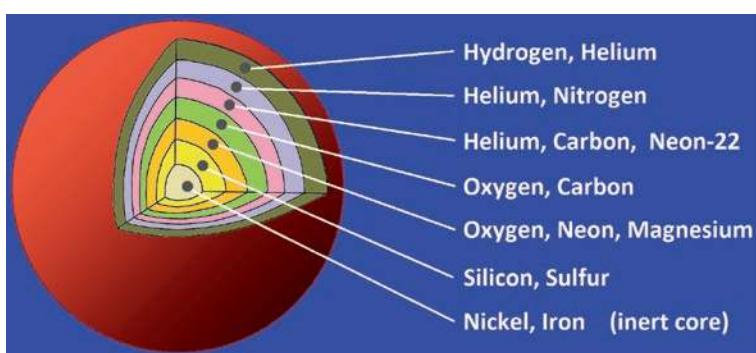


Fig. 13.5 Stellar nucleosynthesis. The heaviest element that can be synthesized in the core is iron. Other heavy elements are burnt in shell-like layers close to the center, while lighter elements continue to burn in the outer layers, where the temperature is lower

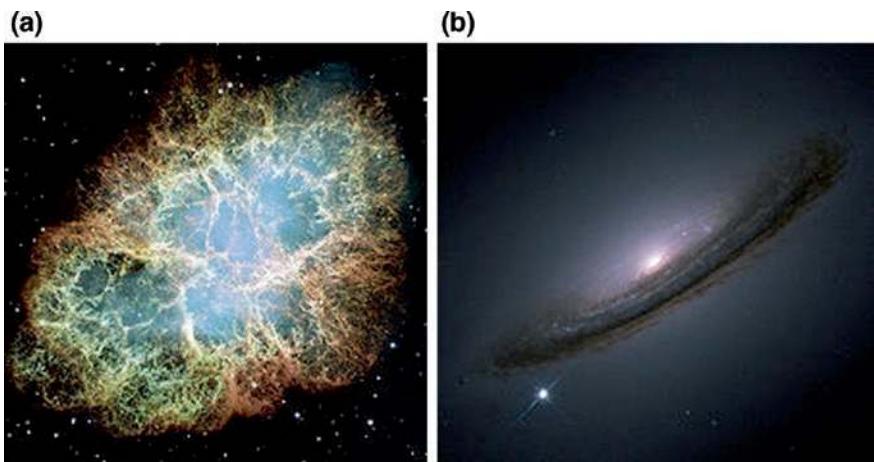


Fig. 13.6 **a** The Crab nebula is the aftermath of a supernova explosion in our galaxy, recorded in AD 1054. Credit STSCI, NASA, ESA, J. Hester and A. Loll (Arizona State University). **b** The bright spot on the lower left is a supernova that competes in brightness with the entire host galaxy. Credit NASA/ESA, The Hubble Key Project Team and The High-Z Supernova Search Team

Elements heavier than iron are forged during core collapse and the gigantic supernova explosion which immediately follows⁴. These heavy elements are expelled into the interstellar medium where they serve as raw material for new stars and planets.⁵ Planets form as a natural by-product of star formation. Thus in a very real sense, we are recycled star dust—the carbon in our cells, the iron in our blood and the calcium in our bones were all made in the centers of stars, and then recycled into the universe (Fig. 13.6).

13.4 Planetary System Formation

All planetary systems, including our Solar System, are believed to have formed in about the same way. A large, slowly rotating cloud of gas begins to contract due to gravitational forces. As the cloud contracts the rotation speeds up, much like an ice-skater spins faster as her hands are pulled in

⁴If the progenitor star has a mass between 8 and 20 Solar masses, the remnant left after core collapse is a super-dense neutron star. Core collapse in stars more massive than 20 Solar masses is expected to produce black holes.

⁵We emphasize that primordial nucleosynthesis is the only explanation we have for the abundances of helium and other light elements. This particularly applies to deuterium, which can only be destroyed in stars. And the amount of helium produced in all stars is only around one percent of the total amount observed in the universe.

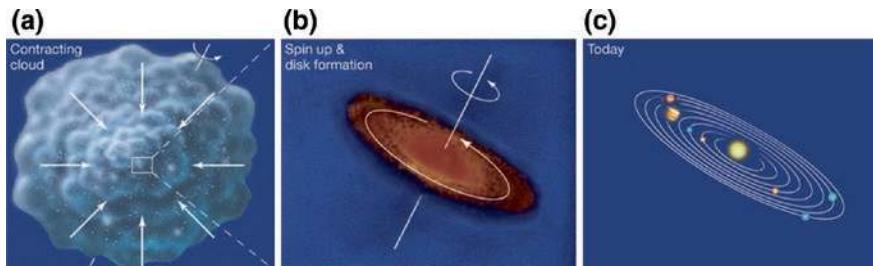


Fig. 13.7 Formation of a planetary system. **a** A rotating cloud of gas starts to collapse. **b** The collapsing material collects mostly near the center and flattens into a rotating disc at the periphery. The central material becomes a protostar and some clumps of matter start to stick together in the disk. **c** Wind from the protostar clears out most of the surrounding material. The remaining material continues to stick together to form local aggregates of matter. Planets which orbit in the same direction and in the same plane are eventually formed. Credit Eric Chaisson (from Astronomy Today, Eric Chaisson, Stephen McMillan, Columbus (Ohio))

towards her body.⁶ The combined action of gravity and rotation causes the cloud to flatten into a thin disc. As the cloud contracts, the material becomes denser and hotter, especially towards the center, and eventually the central region becomes a star, and some of the material in the disk coalesces into a series of planets (see Fig. 13.7).

Our Solar System consists of the Sun, eight planets, an asteroid belt between Mars and Jupiter, and the Kuiper belt beyond Neptune's orbit (see Fig. 13.8). Almost all of the mass in the Solar System is concentrated in the Sun, with most of the remaining mass in the largest planet—Jupiter. The first four planets (Mercury, Venus, Earth and Mars) are called terrestrial planets as they have the same rocky makeup as the Earth; the next four planets (Jupiter, Saturn, Uranus, and Neptune) are the gaseous planets. The asteroid and Kuiper belts consist mostly of bodies that are much smaller than the planets. Jupiter's gravitational bullying prevented the material in the asteroid belt from becoming a planet, and the material in the Kuiper belt probably never collided frequently enough to coalesce into a planet.

Planets outside of the Solar System are very hard to detect, because light reflected by any planet is very dim compared to the brightness of the star it orbits. Astronomers therefore use indirect detection methods, looking for minute effects that the orbiting planet has on the spectrum and the brightness of the star. The first successful detection of an extrasolar planet was

⁶The physical principle behind the spin-up of a contracting cloud (and the ice-skater) is angular momentum conservation.

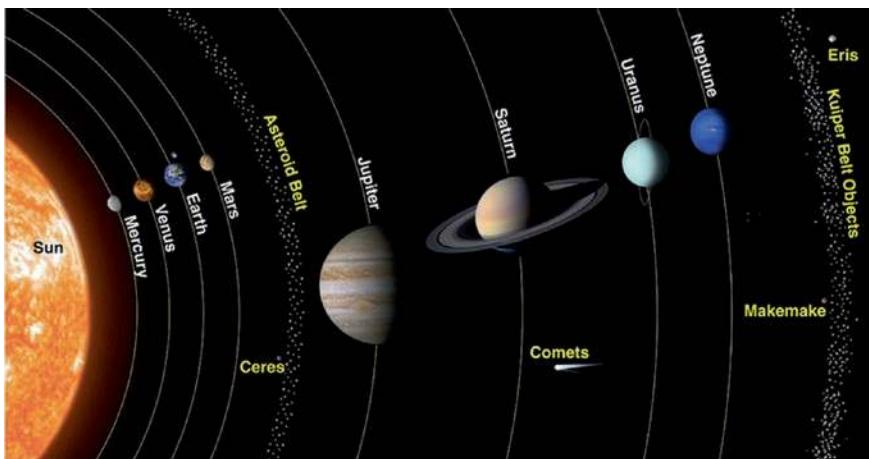


Fig. 13.8 The Solar System (not to scale). Credit NASA

made in 1992 using the Doppler shift method. The gravitational pull of the planet causes the star to move ever so slightly, inducing a Doppler shift in the spectrum of the star. The spectrum is periodically blue and red-shifted as the star moves forward and backward relative to the Earth. This method works best for very massive planets located in close proximity to their stars.

An alternative method is to measure how much a star dims when a planet passes in front of it. The amount of dimming is proportional to the fraction of the stellar disc blocked and grows with the size of the planet. By measuring the duration of each dimming and the time interval between them, one can also determine the revolution period and the radius of the planet's orbit. The Kepler space telescope, launched in 2009, used this method to discover thousands of extrasolar planets, indicating that planet formation is quite common in the universe. On average, the estimated number is more than one planet per star.

13.5 Life in the Universe

With a multitude of planetary systems already discovered, one cannot help but wonder how many of them might harbor life and intelligence, and what exotic forms that alien life might take. Life as we know it on Earth is based on chains of carbon atoms. This is probably not an accident: carbon is one of the most abundant elements and it has the richest chemistry. Other elements involved in life—hydrogen, oxygen and nitrogen—are also among

the most abundant in the universe (see Fig. 13.2). Another key ingredient necessary to enable biochemical reactions is liquid water, which plays the role of a solvent.⁷

There is no shortage of water on most extrasolar planets, but it is either frozen or vaporized if the planet is too far or too close to the star. Astronomers estimate that about 20% of stars have Earth-sized planets with surface temperatures allowing for the existence of liquid water. With about 10^{22} stars in the observable universe, this amounts to $N_p \sim 10^{21}$ potentially habitable planets.

This is a lot of planets, but the number of planets that are actually inhabited by living creatures is very hard to assess. Living organisms are characterized by their ability to reproduce and to undergo Darwinian evolution. Once evolution starts, species proliferate and adapt to the changing environment. This makes life very resilient: life on Earth has survived a number of catastrophic climate changes induced by asteroid impacts, massive volcano eruptions, etc. But in order to get the evolution going, the first living organism had to be formed somehow. How this could happen is presently a great mystery and a subject of ongoing scientific debate.

One can imagine that life started with a relatively short molecular chain, which could replicate itself. But the probability for such a chain to arise by chance from non-living matter is likely to be very small. Another road-block on the way to life is the origin of the genetic code and of the complex molecular machinery that builds proteins (of which all living organisms are made) following the instructions encoded in DNA. At the present level of understanding, we cannot exclude the possibility that fluctuations necessary for the evolution of life are so rare that life on Earth is the only life that has ever evolved within our cosmic horizon.

Fossil evidence indicates that microbial life existed on Earth about 3.8 billion years ago. Prior to that, the Earth was heavily bombarded by asteroids and life was probably impossible. Thus, it seems like life emerged almost as soon as it could. At first glance, this suggests that life appears easily, and therefore the universe should be teeming with primitive life forms. But it could also be that the rare fluctuation necessary for the emergence of life was more likely to occur in the turbulent environment of the early Earth. Then the most probable time to form life would be soon after a planet forms, but only a small fraction of planets would have life.

⁷Alternative forms of biochemistry have also been hypothesized. For example, life might be based on silicon and use ammonia as a solvent. Here we focus on Earth-like life.

Even if primitive life is abundant in the universe, the chances for it to evolve intelligence are highly uncertain. Dinosaurs roamed the Earth for more than 150 million years and did not develop a technological civilization. We may soon have a better idea of how widespread intelligent life actually is. An extensive international program of observations searching for signs of extraterrestrial intelligent life, called *Breakthrough Listen*, was launched in 2016.

Summary

The lightest atomic nuclei—hydrogen, helium, and small amounts of deuterium and lithium—were formed during the first few minutes after the big bang. The theory of big bang nucleosynthesis tightly constrains the amount of atomic matter that can exist in the universe and leads to the conclusion that most of the dark matter consists of some unknown particles.

Elements up to iron are produced in the cores of stars, and heavier elements are made during violent supernova explosions, which then spew these heavy elements into the interstellar medium. New generations of stars form from the enriched interstellar gas, and planets form as by-products.

We now know that planetary systems are abundant in the universe, and astronomers estimate that there are a huge number of habitable planets. We still do not know how to estimate the probability for primitive life to arise on a generic habitable planet. Even if primitive life is common, intelligent life may still be relatively rare. If so, we may be the lone technologically advanced civilization in our cosmic horizon.

Questions

1. What determines the chemical properties of an atom?
2. If a carbon nucleus has six protons, how many electrons does a carbon atom have?
3. Why were alchemists unable to produce gold from other elements?
4. What are the two most abundant elements in the universe? When were most of these two elements formed?
5. Roughly what percentage (by mass) of atomic matter in the universe is in the form of hydrogen? And helium?
6. Where were elements heavier than lithium formed?
7. At roughly what age did the universe complete big bang nucleosynthesis?
8. The Sun consists mostly of hydrogen. What happens to the hydrogen in the central parts of the Sun?

9. Why do fusion reactions in stars need such high temperatures? Why does formation of increasingly heavy nuclei require increasingly higher temperatures?
10. How do we know what chemical elements exist in stars and interstellar gas?
11. Take note of the different elements in your surroundings, and try to imagine where these elements were created, and how they made their long journey from then to now.
12. Why did element formation stop with helium fusion in the early universe?
13. Cosmologists believe that most of the dark matter cannot be ordinary atomic matter. Can you explain how this conclusion follows from the theory of big bang nucleosynthesis?
14. How do the primordial deuterium and lithium abundances allow us to determine the current total density of atomic matter?
15. In massive stars different types of nuclear reactions take place creating a variety of different elements. Where in the star are the heaviest elements made and why?
16. When a massive star runs out of nuclear fuel at its center, what happens to the central core? What happens to the outer layers of the star?
17. Name one reason why supernovae are critical for your existence.
18. Why does a protostellar cloud heat up as it contracts?
19. A distant star, having a brightness and a radius very similar to our Sun, is observed to get dimmer by about 0.01% every 400 days, with every dimming episode lasting 12 h. Assuming that the dimmings are caused by an orbiting planet, estimate the radius of the planet, the speed of its revolution around the star, and the radius of its orbit. (Note: the radius of the Sun is 7×10^8 m.)
20. Astronomers call a planet “habitable” if it has liquid water. Does this mean the planet is actually inhabited by some living creatures? Would you expect that most “habitable” planets harbor some forms of life?

14

The Very Early Universe

At the time of nucleosynthesis, the primordial fireball consisted of electrons, protons, neutrons, photons and neutrinos. As we get closer to the big bang, the fireball gets hotter and denser, and other types of particles emerge. They move at nearly the speed of light, colliding frequently and violently. As we shall see, some of the most dramatic events in the history of the universe occurred within a small fraction of a second after the big bang.

14.1 Particle Physics and the Big Bang

The density and temperature of the universe increase as we follow their evolution backward in time, towards the big bang. If time is measured in seconds, then the density and temperature are given by¹

$$\rho \approx \frac{4.5 \times 10^8}{t^2} \text{ kg/m}^3 \quad (14.1)$$

and

$$T \approx \frac{10^{10}}{\sqrt{t}} \text{ K} \quad (14.2)$$

¹These relations hold during the *radiation era*. While we won't derive these equations here (see the Appendix), we will outline how the dependence on time emerges. The energy density is proportional to the inverse fourth power of the scale factor, $\rho \propto a(t)^{-4}$; the temperature scales as the inverse scale factor, $T \propto a(t)^{-1}$; and the scale factor is proportional to the square root of time, $a(t) \propto \sqrt{t}$ (found by solving Friedman's equation during the radiation era). Thus, $\rho \propto t^{-2}$, and $T \propto t^{-\frac{1}{2}}$.

The temperature of the early universe is proportional to the average photon energy. Physicists measure particle energies and masses in electron-volts. One electron-volt is the energy gained by an electron as it moves across a potential difference of one volt, $1\text{ eV} = 1.6 \times 10^{-19}\text{ J}$; the equivalent mass is $1\text{ eV} = 2 \times 10^{-36}\text{ kg}$. Other related units are the $\text{MeV} = 10^6\text{ eV}$, $\text{GeV} = 10^9\text{ eV}$, and $\text{TeV} = 10^{12}\text{ eV}$. If energy is measured in MeV and temperature in Kelvins, the average energy per photon is roughly

$$E \sim 10^{-10}T\text{ MeV}. \quad (14.3)$$

Thus, an energy of $E = 1\text{ MeV}$ corresponds to a temperature of $T \sim 10^{10}\text{ K}$.

It follows from Eqs. (14.1) and (14.2) that at about 100 s (when nucleosynthesis takes place), the density of the universe is about 50 times that of water ($\rho = 50,000\text{ kg/m}^3$), and the temperature is a billion Kelvin ($T = 10^9\text{ K}$), which is about 100 times hotter than the core of the Sun. At 1 s, the density is 500 thousand times that of water, and the temperature is 10 billion Kelvin. Jumping back to a microsecond, the density soars to $\rho \sim 10^{21}\text{ kg/m}^3$, and the temperature is ten trillion Kelvin ($T = 10^{13}\text{ K}$). The primordial cauldron is an extreme environment! The closer we get to the big bang, the more energetic the particles become and the more frequently they smash into one another. It is therefore important to understand what happens in such high-energy collisions.

Physicists study particle collisions with the help of colossal machines called accelerators. Inside an accelerator particle accelerator, particles are boosted to extremely high energies by electric fields, and then oppositely directed particle beams are aimed at one another in a small area surrounded by detectors. By studying the collision debris, physicists try to figure out the laws governing high-energy particle interactions (Fig. 14.1).

In a high-energy particle collision, there are generally a number of possible outcomes, which occur with different probabilities. The range of possibilities is restricted by a few *conservation laws*, such as energy and charge conservation: the total energy and the total electric charge should be the same before and after the collision. Other conserved quantities include baryon and lepton numbers and the “color” charge. Any process that is not forbidden by conservation laws will occur with some nonzero probability, which can be calculated using the rules of quantum mechanics.

A remarkable property of particle encounters is that the colliding particles can change their identity. For example, a pair of photons can turn into an electron-positron pair, as illustrated in Fig. 14.2. The positron is the electron’s antiparticle—all particles have an antiparticle with identical properties, except for having opposite charges. Photons are their own antiparticles;



Fig. 14.1 The Large Hadron Collider (LHC) in Geneva Switzerland lies underground in a circular tunnel with a circumference of nearly 30 km. It is the largest particle accelerator in the world. The LHC achieves energies of $E = 14 \text{ TeV}$, which correspond to temperatures of $T = 10^{17} \text{ K}$ and a time of $t = 10^{-14} \text{ s}$ after the big bang Credit CERN

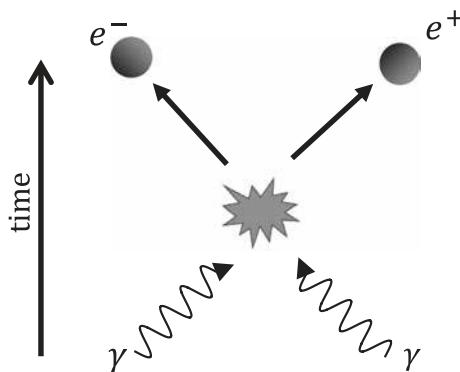


Fig. 14.2 Two photons collide to produce an electron-positron pair

their electric and other charges are all equal to zero. Particles and antiparticles are often created in pairs, as in the electron-positron pair production. The inverse process, called pair annihilation occurs when a particle and an antiparticle collide and turn into two photons. The number of particles can also be changed: two initial particles can produce a hail of other particles flying away from the collision point (see Fig. 14.3). These types of events were commonplace in the early moments after the big bang.

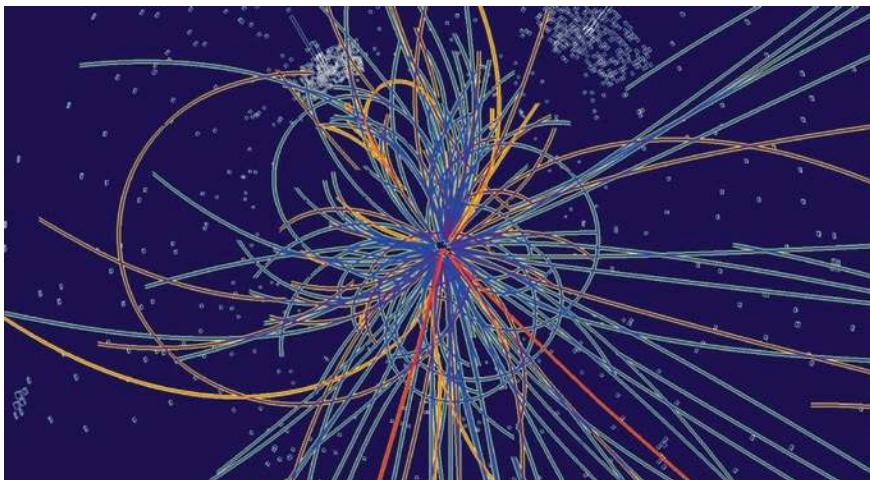


Fig. 14.3 A multitude of particles is produced in a proton-proton collision. The trajectories of positively and negatively charged particles are curved by magnetic fields, resulting in the circular paths

Energy conservation requires that a pair of photons creating a particle-antiparticle pair must have a combined energy of at least $2 mc^2$, where m is the particle's mass.² The electron mass is $m_e \approx 0.5$ MeV, so electron-positron pairs are copiously produced at temperatures greater than 5×10^9 K (corresponding to photon energies $E \gtrsim 0.5$ MeV). As a result, the primordial fireball gets populated with electrons and positrons having about the same density as photons, and about the same energy per particle. Under these conditions, pair creation by photon collisions occurs at the same rate as pair annihilation due to electron-positron encounters, so electrons, positrons and photons are in thermal equilibrium.

At still higher temperatures, more massive particles and antiparticles appear. Muons, for example, have a mass of $m_\mu \approx 100$ MeV; in the first microsecond they are abundantly present in equilibrium with their antiparticles, at $T > 10^{12}$ K. Each kind of particle has a threshold temperature that must be reached in order for that particle type to be created in large numbers. In the early universe, as the fireball expands and cools, particles and antiparticles annihilate, and cannot be replenished, when the temperature drops below the corresponding threshold. Thus, muons annihilate with anti-muons at $t \sim 10^{-6}$ s, and electron-positron pairs annihilate at $t \sim 1$ s ABB.

²When we say a particle has mass m , we usually refer to its rest mass.

14.2 The Standard Model of Particle Physics

Particle physicists have developed a theory, called the Standard Model, which accurately describes most of the known particles and their interactions (see Fig. 14.4). Particles can be divided into *matter* particles called fermions, and force particles, called *gauge bosons*. In addition, there is a particle called the Higgs boson, which plays a special role in the theory, as we shall explain below.

Technically, the classification of particles into fermions and bosons is based on a quantum property called spin. Very roughly, a particle can be pictured as a tiny ball rotating about its axis, with spin characterizing the intensity of rotation. Spin can take only a discrete set of values: 0, 1/2, 1, 3/2, Fermions have half-integer spin, and bosons have integer spin. All fermions of the Standard Model have spin 1/2, all gauge bosons have spin 1, and the Higgs boson has spin 0.

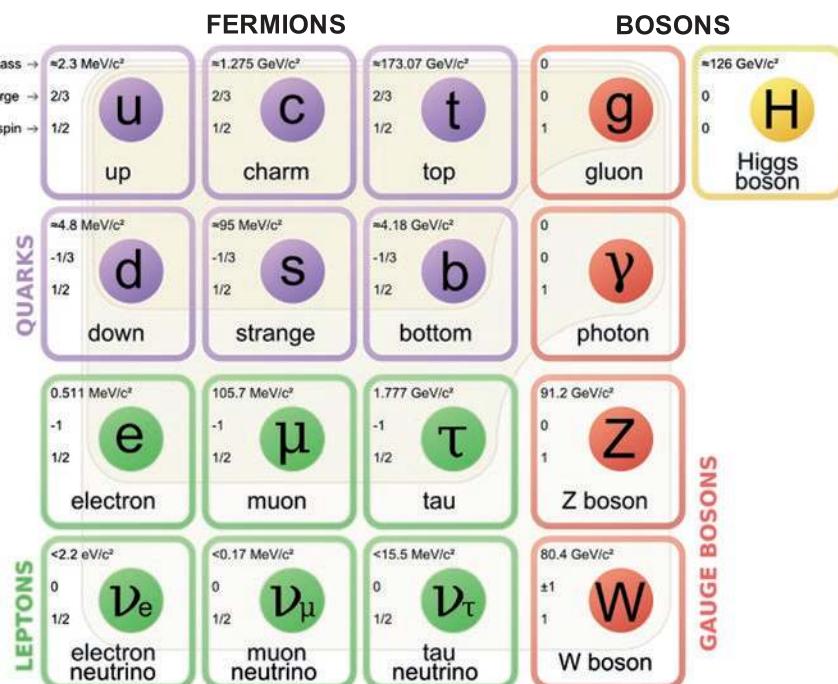


Fig. 14.4 The particles of the Standard Model, including their mass, electric charge and spin

14.2.1 The Particles

The elementary fermions include six quarks (with the whimsical names³: up, down, charm, strange, top and bottom) and six leptons (the electron, muon, tau particle, and their associated neutrinos). Individual quarks are never observed in nature; they are always tightly bound together into composite particles called hadrons. The most familiar hadrons are protons (composed of two up and one down quarks) and neutrons (composed of one up and two down quarks). All other hadrons are unstable. They can be produced in particle accelerators, but thereafter decay in a small fraction of a second. Leptons do not bind together like quarks do. All three neutrinos are stable, very light, and interact extremely weakly. The muon and the tau have the same electric charge as the electron, but are much heavier and rapidly decay into electrons and neutrinos.

To describe almost all known matter, we need essentially only four particles: the up and down quarks, the electron and the electron neutrino. Together they make up the so-called first generation of elementary particles. Apart from accelerator experiments and some extreme astrophysical processes, the properties of our world would not change if the other two generations (the second and third columns in Fig. 14.4) did not exist.

14.2.2 The Forces

All the interactions between particles can be described by four forces: gravity, electromagnetism, the weak nuclear force and the strong nuclear force. Although we are all very familiar with gravity, it is the only force (that we know of) that is not described by the Standard Model of particle physics—we will return to this important distinction later. All particles interact gravitationally via the hypothesized graviton, which has yet to be observed.

The electromagnetic force acts between electrically charged particles. It is mediated by photons: one particle emits a photon and the other absorbs it, as shown in Fig. 14.5. This force gives rise to most of the physics we are familiar with, and to all of chemistry. It is the glue that keeps electrons in atoms and is responsible for interactions between atoms and molecules.

As its name suggests, the strong nuclear force is the strongest of the four interactions. It binds quarks into nucleons and holds nucleons together

³The quark names have no meaning other than serving to distinguish the different quark particles. For example, “up” and “down” have nothing to do with direction.

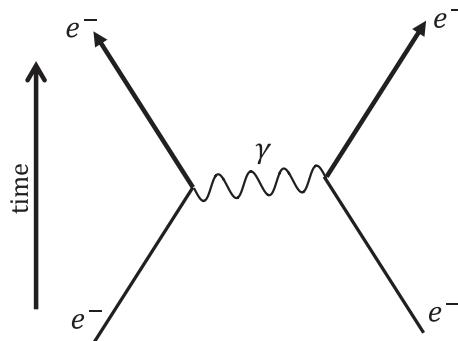


Fig. 14.5 Two electrons are electromagnetically repelled as they exchange a photon

in nuclei. Quarks carry the so-called “color” charge and interact through exchange of gluons, in much the same way as electrically charged particles interact through the exchange of photons. An important difference is that photons carry no electric charge while gluons themselves carry a color charge and can emit and absorb other gluons.

The weak nuclear force is responsible for some radioactive decays and for the interactions of neutrinos. It is mediated by the massive W and Z bosons. The weakness of this interaction and its short (microscopic) range are due to the large masses of its gauge bosons.

The Higgs boson does not mediate any force, but this particle and its associated field, called the Higgs field, play a key role in the Standard Model. The Higgs field is related to the Higgs particle in the same way as the electromagnetic field is related to the photon. The Higgs boson has a large mass and a short lifetime; its production requires very powerful accelerators. It was finally discovered in 2012, almost 50 years after its prediction in 1964 by Peter Higgs and by Francois Englert and Robert Brout. The Higgs field is present all around us, at every point in space, and its presence has a dramatic impact on the physical character of our world.

To illustrate the properties and the significance of the Higgs field, we can compare it to the magnetic field, which also permeates the space around us. We cannot feel the magnetic field with our sense organs, but we can detect its presence using a compass or by observing trajectories of charged particles;

instead of a straight line, a particle in a magnetic field moves along a spiral path.⁴ Magnetic fields are produced by electric currents; for example, the magnetic field of the Earth is due to the currents flowing in the core of our planet. But as one gets very far away from planets and stars, the magnetic field strength declines towards zero.

The Higgs field, on the other hand, is non-zero even in vacuum; it has the same strength everywhere in the universe. Another distinction is that the magnetic field is a vector—it is characterized by both magnitude and direction. The Higgs field is characterized only by its magnitude. Such fields are called *scalar fields*. In this regard the Higgs field is similar to the temperature, which also has some magnitude at each point in space and time, but no direction.

If we could vary the magnitude of the Higgs field, we would feel the unpleasant effects immediately. The masses of all matter particles in the Standard Model (except neutrinos) are proportional to the Higgs field. So, if the strength of the field were changed, the masses would also be altered, resulting in some new physical and chemical properties of all matter. If we could turn off the Higgs field completely, all Standard Model particles would become massless and would move at the speed of light. In particular, the W and Z bosons would be massless, like photons, and the weak nuclear force would be essentially indistinguishable from electromagnetism. The universe would be a very different place! So, why is the Higgs field non-zero?

14.3 Symmetry Breaking

There is a good reason why the magnetic field is zero in vacuum. Fields, like particles, have a certain amount of energy. In any region where the magnetic field is non-zero, it has an energy density proportional to the square of the field; see Fig. 14.6. As the field is increased, its energy grows. Since the vacuum is the state of lowest energy, the magnetic field must vanish in vacuum.

The energy density of the Higgs field exhibits a very different behavior, which is schematically illustrated in Fig. 14.7. The lowest energy states are now at non-zero values of the field, which are labeled V and $-V$ in the figure. They are the vacuum states of the theory. It does not matter which of

⁴The spiral curves in opposite directions for positively and negatively charged particles, and its radius depends on the particle's energy. Physicists use these properties to analyze high-energy collisions, like the one shown in Fig. 14.3.

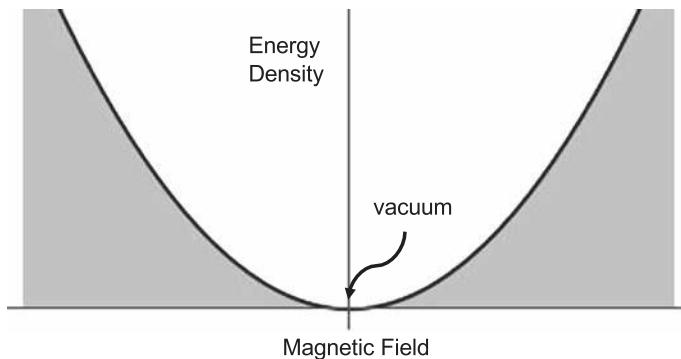


Fig. 14.6 Energy density of a magnetic field vs the magnitude of the field. When the field is zero, the energy density is also zero

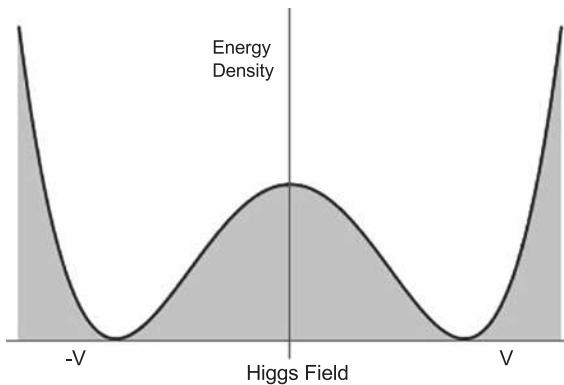


Fig. 14.7 Higgs field potential energy density. There are two vacuum states which have non-zero Higgs field values, labelled $-V$, and V

these states nature chooses: they have identical physical properties. But with either choice the Higgs field is non-zero in the vacuum. The two vacuum states are separated by a high energy hill, with the top of the hill corresponding to vanishing Higgs field. Setting the Higgs field to zero, even in a small region of space, would thus be a very costly proposition. For just one cubic centimeter the required energy far exceeds the present energy resources of our planet.

The energy dependence of the kind shown in Fig. 14.7 is not uncommon in nature and can even occur for an ordinary magnetic field. The case in point is a simple bar magnet (Fig. 14.8). Each atom in a magnet acts like a tiny magnet itself. Interaction between these microscopic magnets causes them to align, resulting in a large magnetic field. The energy curve for a bar

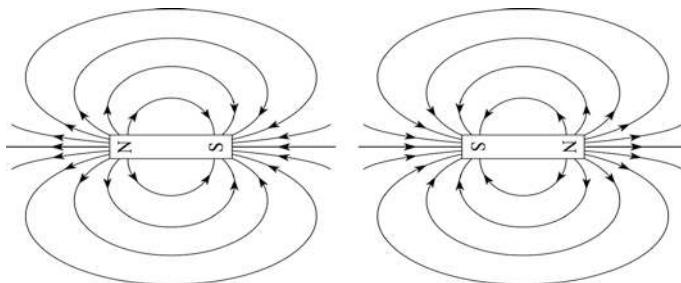


Fig. 14.8 The magnetic field of a bar magnet (*left*) and its rotated version about the vertical axis (*right*). The direction of magnetization reverses, but the energy is the same

magnet as a function of the magnetic field has the same form as in Fig. 14.7. The two energy minima now correspond to the same strength of the field, but opposite directions of magnetization. Clearly, these two states have identical properties, since they can be obtained from one another by rotating the magnet 180° about a vertical axis. In the case of a magnet, the energy cost of setting the magnetic field to zero is not prohibitively high. When the magnet is heated to a temperature above 10^3 K (the so-called Curie temperature), the alignment of atoms is destroyed by random thermal motions, and macroscopic magnetization disappears. If the magnet is then cooled below the Curie point, it gets spontaneously magnetized, with the direction of magnetization selected randomly by thermal fluctuations.⁵

We now introduce the important concepts of symmetry and symmetry breaking. We say that a physical system has symmetry if there are some transformations that leave it unchanged. For example, a spherical object is symmetric, since it does not change if we rotate it about any axis passing through its center. An iron bar heated above the Curie temperature has symmetry with respect to flipping. But the same bar below the Curie temperature has no such symmetry. It gets magnetized, and the magnetization direction is reversed when the bar is flipped. We say in such cases that the symmetry has been broken.

Getting back to the Standard Model, the state with a vanishing Higgs field has a high degree of symmetry. Matter particles in this state are

⁵This picture of spontaneous magnetization applies only to very small magnets. When a large piece of iron is cooled below the Curie temperature, it splits into a number of domains with different directions of magnetization.

interchangeable, since they all have zero mass.⁶ The weak nuclear force is indistinguishable from electromagnetism; together, they are referred to as the electroweak force, and the symmetry between them is called the *electroweak symmetry*. As we discussed, however, this symmetric state does not correspond to the minimum of energy, so the Higgs field becomes non-zero and the symmetry gets broken.

Just like in a bar magnet, the electroweak symmetry is restored at sufficiently high temperatures. But in this case the universe has to be heated above 10^{15} K! The average particle energies are then $E > 100$ GeV, high enough to populate the fireball with W , Z and Higgs bosons. These extreme conditions are recreated in high-energy particle collisions. As predicted, experiments show that the differences between the weak and electromagnetic forces disappear at energies above 100 GeV.

14.4 The Early Universe Timeline

We are now ready to summarize the important milestones of the early universe.

Electroweak symmetry breaking: $t \sim 10^{-10}$ s ($T \sim 10^{15}$ K)

Before this event, all Standard Model particles are massless; they (and their antiparticles) populate the fireball with about the same density as photons. As the symmetry gets broken, all particle masses become different, and the weak and electromagnetic interactions become distinct. Soon after that, W , Z and Higgs bosons annihilate with their antiparticles.

Quark confinement: $t \sim 10^{-6}$ s ($T \sim 10^{13}$ K)

We mentioned in Sect. 14.2 that individual quarks are never observed; they are confined by gluons into protons and neutrons. However, at times earlier than a microsecond ABB, the density of protons, neutrons and their antiparticles is so high that they overlap, and their constituent quarks mix together, forming a dense gas of quarks, antiquarks and gluons. At $t \sim 10^{-6}$ s, the density of this gas drops enough for quarks to become bound together into non-overlapping protons and neutrons. The average particle energy at that time is $E \sim 1$ GeV. Almost all the protons, neutrons, and their antiparticles annihilate soon thereafter, producing photons (and also other light particles,

⁶More precisely, leptons cannot be distinguished from other leptons and quarks from other quarks, but quarks can be distinguished from leptons, since only quarks can interact through the exchange of gluons.

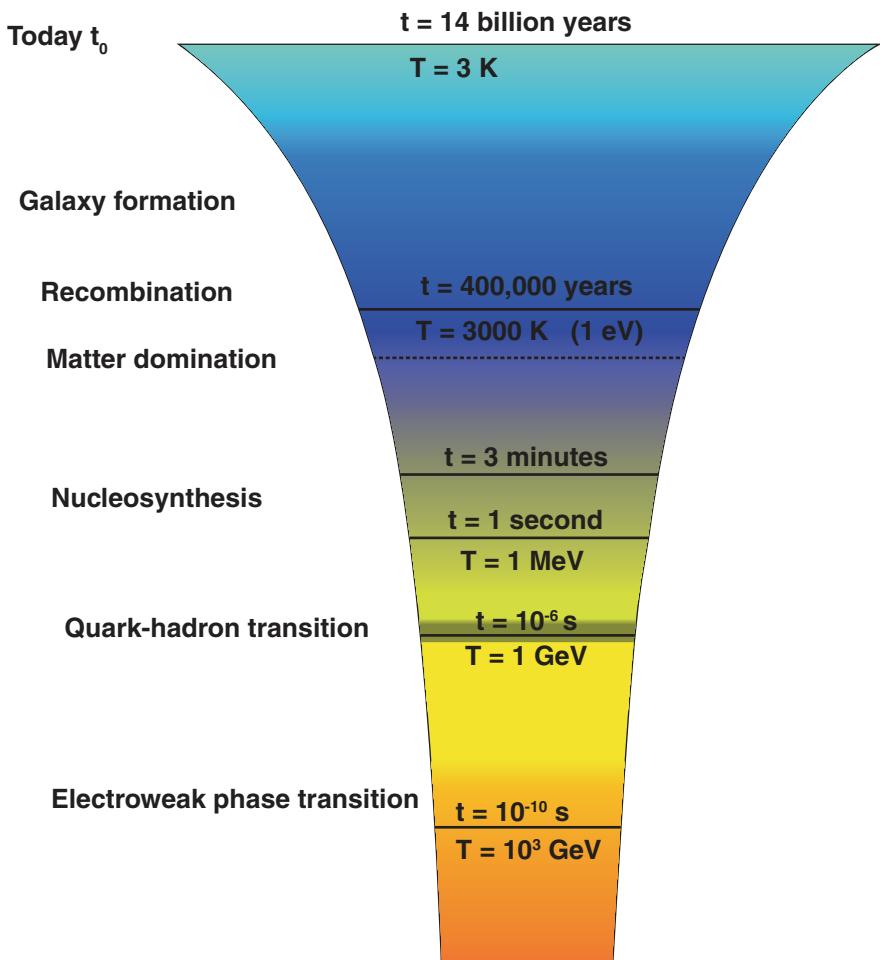


Fig. 14.9 Early universe timeline

like electrons, muons and neutrinos). But they could not all annihilate, since otherwise there would be no nucleons left to form atomic matter. This tells us that there was a small imbalance: nucleons outnumbered antinucleons by about one in a billion, so that the nucleon to photon ratio after the annihilation is $\frac{n_n}{n_\gamma} \sim 10^{-9}$. The surviving nucleons thus become the first bound systems that emerge as the universe cools.

Electron-positron annihilation: $t \sim 1 \text{ s}$ ($T \sim 10^{10} \text{ K}$)

Between 10^{-6} s and 1 s ABB, all remaining particle-antiparticle pairs disappear from the fireball. The last to disappear are the electron-positron pairs, which annihilate at $t \sim 1 \text{ s}$. Like the case of nucleons and anti-nucleons,

there must have been a small excess ($\sim 10^{-9}$) of electrons over positrons, so that atomic matter could later be formed.

Nucleosynthesis: $t \sim 100$ s ($T \sim 10^9$ K)

Protons combine with neutrons that have not yet decayed to form helium and other light nuclei. The remaining protons exist freely as hydrogen nuclei.

Recombination: $t \sim 400,000$ yrs ($T \sim 3000$ K)

Nuclei are bound together with electrons to form neutral atoms. Photons of the fireball are now free to propagate through the neutral atomic gas and reach us in the form of cosmic microwave background radiation.

The timeline of the early universe is illustrated in Fig. 14.9, where we have also included some key events related to structure formation. Note that some of the most important events in the cosmic history occurred within the first second ABB.

14.5 Physics Beyond the Standard Model

The Standard Model describes much of the amazing variety and complexity of our physical world—and yet it is incomplete. It does not account for neutrino masses, and the gravitational force lies outside its scope. Moreover, as we discussed in Chap. 13, dark matter cannot be made of ordinary atoms and must consist of some unknown particles, not included in the Standard Model.

14.5.1 Unifying the Fundamental Forces

An overarching theme in the history of physics has been the idea of unification. It has long been a dream of particle physicists to develop a unified theory that describes all particles and their interactions. Einstein himself spent the last thirty years of his life, struggling (unsuccessfully) to unify electromagnetism with gravity.

In 1864, Maxwell's theory of electromagnetism unified the previously separate phenomena of electricity and magnetism. About 100 years later, scientists developed the *electroweak theory*,⁷ which unified electromagnetism

⁷Steven Weinberg, Abdus Salam and Sheldon Glashow shared the 1979 Physics Nobel prize for this work.

and the weak interactions at energies $E \sim 100$ GeV ($T \sim 10^{15}$ K). The strong interactions are described by a separate theory, called *quantum chromodynamics*.⁸ The Standard Model includes the electroweak theory and quantum chromodynamics as two independent parts. There are a number of candidate theories, collectively called Grand Unified Theories (GUTs), which attempt to unify the electroweak and strong interactions. Analysis shows that the strong nuclear force gets weaker with increasing energy and becomes comparable to the electroweak force at $E \sim 10^{16}$ GeV. Hence the grand unification must occur at very high energies and temperatures ($E \sim 10^{16}$ GeV, $T \sim 10^{29}$ K). Ultimately, gravity also needs to be unified with the other forces. *String theory* is currently the most promising framework to accomplish this goal (see Chap. 19). It suggests that gravity and the GUT force merge at the Planck⁹ energy scale ($E \sim 10^{19}$ GeV, $T \sim 10^{32}$ K).

In the cosmological context, as the universe cools down from the big bang, it goes through a series of symmetry breaking transitions (see Fig. 14.10). Each transition has a Higgs field associated with it; this field is equal to zero prior to the transition and takes a non-zero value once the symmetry is broken. The symmetry breaking transitions occur in rapid succession, so all four interactions become distinct within a small fraction of a second after the big bang.

The electroweak unification and subsequent symmetry breaking are understood theoretically, and well tested experimentally. Unfortunately, the same cannot be said for GUTs. While theorists have proposed many GUTs, it is not so easy to test them because GUT scale energies are inaccessible to particle accelerators. We would need an accelerator that is 3 light years long—almost all the way to Alpha Centauri—to reach GUT scale energies! There is however hope that studies of the early universe might provide an observational window on GUT scale physics. As the Soviet physicist Yakov Zeldovich put it: “The early universe is the poor man’s accelerator”.

⁸Frank Wilczek, David J Gross and H. David Politzer won the 2004 Physics Nobel prize for their work on quantum chromodynamics.

⁹Max Planck pointed out that the only quantity with dimension of energy that can be constructed out of the fundamental constants G , c and \hbar is: $E = \sqrt{\frac{c^5 \hbar}{G}} \sim 10^{19}$ GeV. This is the Planck energy.

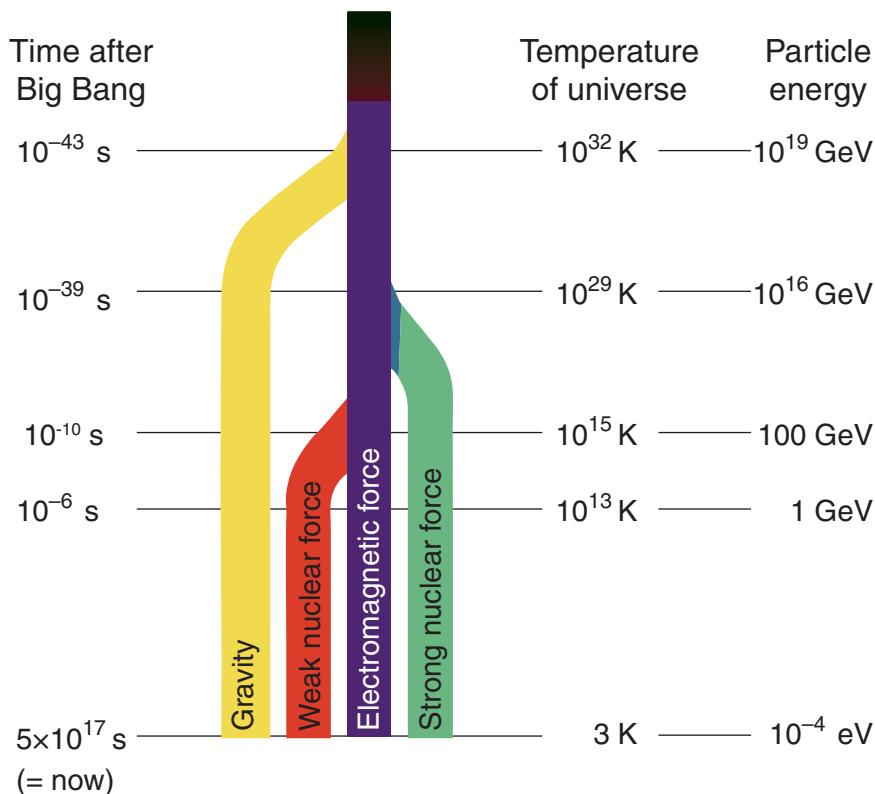


Fig. 14.10 Unification of the four forces. Today the four forces all have a separate identity, but when the universe was much hotter this was not the case. The weak and electromagnetic forces were once the electroweak force. Grand Unified Theories suggest that the electroweak and strong nuclear force were also once united, and split into two forces as the universe underwent symmetry breaking at the GUT energy scale of about 10^{16} GeV . In some models the GUT symmetry is broken in several steps; then there are some additional symmetry breaking transitions between the GUT and electroweak energy scales. String theory seeks to explain how gravity merges with the GUT force at the Planck energy $E \sim 10^{19} \text{ GeV}$ ($T \sim 10^{32} \text{ K}$).

14.6 Vacuum Defects

Symmetry breaking transitions in the early universe can produce a variety of peculiar objects, called vacuum defects, which can still be present in the universe today. Depending on the kind of symmetry breaking, defects can come in three basic types—domain walls, strings, and monopoles. Domain walls are surface-like defects; they are thin sheets of concentrated energy

(and mass). Strings are thread-like, with energy distributed along a line. And monopoles are point-like objects, like particles; their energy is concentrated around a single point. Different GUTs predict different kinds of defects. Thus, observation of vacuum defects could provide valuable information about particle physics at very high energies. We shall now discuss the properties of different defects and their possible observational effects. (Note: this section can be skipped without impacting the understanding of subsequent chapters.)

14.6.1 Domain Walls

The simplest model predicting domain walls is the one illustrated in Fig. 14.7.¹⁰ It has two vacuum states, separated by an energy hill. At very high temperatures the Higgs field is equal to zero. Then, as the universe cools below some critical temperature, the symmetry gets broken, and the field has to take one of the two vacuum values, V or $-V$. The choice between the two vacua is dictated by local random fluctuations, so different parts of the universe end up in different vacua. The universe thus splits into domains having the Higgs field values V and $-V$, as illustrated in Fig. 14.11.

We can estimate the typical size of the domains, which is denoted by the letter ξ in the figure. If the symmetry breaking occurs at cosmic time t , this size cannot exceed the horizon distance $d_{hor} \sim ct$ —simply because no interactions could have occurred over larger distances, so a uniform value of the Higgs field could not be established. For an electroweak-scale symmetry breaking, $t \sim 10^{-10}$ s and the domains must be smaller than 3 cm. For a GUT symmetry breaking, the domains are smaller still, by many orders of magnitude.

Now imagine moving from a domain with the Higgs value V to one with the Higgs value $-V$. By continuity, at the boundary between the domains, the Higgs field has to go through zero. But we know that the energy density of the Higgs field gets large when the field is set to zero. Hence, the boundaries between positive and negative Higgs domains must carry a large energy. To minimize the energy cost, the regions where the Higgs field is close to

¹⁰We introduced the two-vacuum model of Fig. 14.7 to illustrate the Standard Model of particle physics, but we emphasize that this is just a schematic illustration. The Higgs field of the Standard Model has three independent components, and the vacuum structure is more complicated. In fact, the Standard Model does not predict any vacuum defects. If any defects are formed, they are likely to come from higher-energy symmetry breakings.

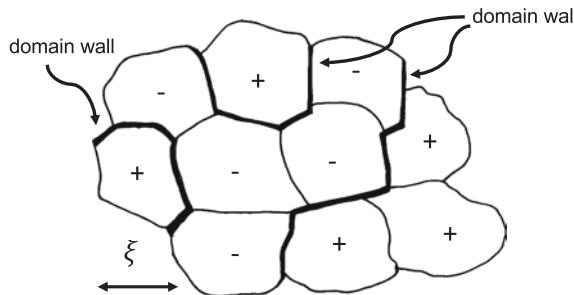


Fig. 14.11 Different physical regions randomly land in one of the two possible vacuum states, separated by domain walls

zero shrink to microscopically thin sheets near the boundaries. These sheets are the domain walls.

You can easily convince yourself that domain walls do not have edges: they can either be closed, completely enclosing domains of positive or negative Higgs field, or they can extend to infinity. The mass per unit wall area depends on the energy scale E_{sb} of symmetry breaking. It is $\sim 10^9 \text{ kg/m}^2$ for the electroweak scale ($E_{sb} \sim 100 \text{ GeV}$) and grows proportionally to E_{sb}^3 at higher energies. Thus, for a GUT-scale symmetry breaking the walls carry a mammoth mass of 10^{51} kg/m^2 .

Once the symmetry is broken, domains with the same value of the Higgs field begin to merge and grow larger. But they cannot grow faster than the speed of light, and thus the typical domain size will always remain smaller than the horizon. Applied to the present time, this implies that there should be at least one domain wall stretching across the presently observable region. Such a wall would have mass much greater than the combined mass of all matter in this region. The gravity of the wall would then drastically disrupt the observed isotropy of the galaxy distribution and of the cosmic microwave background. Since no major disruptions of isotropy are observed, it follows that particle physics models predicting domain walls should be ruled out. Thus, even though we have not yet observed any vacuum defects, we have already learned something about high-energy particle physics.

14.6.2 Cosmic Strings

Thread-like string defects are predicted in a wide variety of particle physics models. Here is a summary of their basic properties.

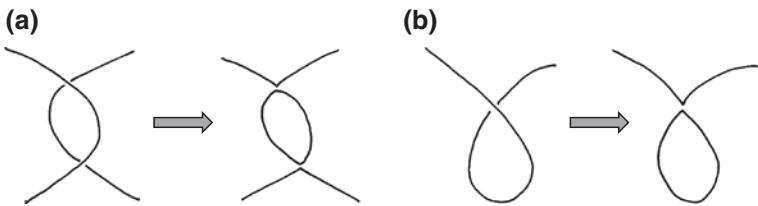


Fig. 14.12 Cosmic strings reconnect as they cross

- Strings do not have ends: they either form closed loops or extend to infinity.
- The thickness of strings is microscopic, while their length can be arbitrarily large. So strings can be well approximated by infinitely thin lines.
- The mass per unit length of a string, usually denoted by μ , is determined by the symmetry breaking energy scale, $\mu \propto E_{sb}^2$. Electroweak-scale strings, if formed, would be very light, $\mu \sim 10^{-7}$ kg/m, while GUT-scale strings would be extremely massive, with $\mu \sim 10^{21}$ kg/m.
- Cosmic strings have a large tension, like in a stretched rubber band. This causes a closed loop of string to oscillate at a speed close to the speed of light.
- When strings intersect, they reconnect, resulting in the formation of closed loops (see Fig. 14.12).

At the time of symmetry breaking, strings form a dense random network, consisting of long, wiggly strings and small closed loops. As for domain walls, the typical distance between the strings in the network cannot exceed the horizon. The subsequent evolution of cosmic strings is rich in physical processes. Tension in wiggly strings causes them to move at relativistic speeds. Moving strings intersect and chop off their wiggles in the form of closed loops. As a result, long strings get straighter with time. Closed loops oscillate and lose their energy by emitting gravitational waves. They gradually shrink and disappear.

Computer simulations have revealed that cosmic string evolution has a scaling property: at any time the string network looks more or less the same, except the overall scale grows proportionally to time. So, if you took a snapshot of the network, say, at $t = 1$ s and magnified it 100 times, it would look very similar to a snapshot taken at $t = 100$ s. In particular, a horizon region at any time contains several long strings and a large number of closed loops, as shown in Fig. 14.13. This applies to the present time as well: if cosmic strings exist, there should be a few long strings stretching across our

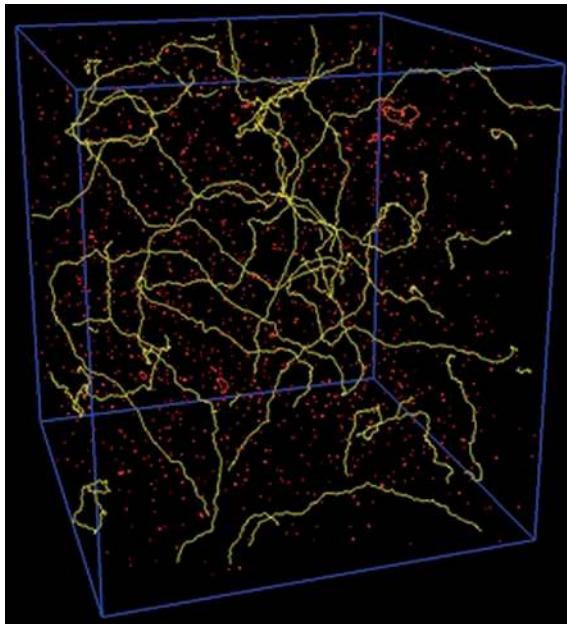


Fig. 14.13 A horizon-size region containing several long strings (shown in yellow) and a large number of small closed loops (shown in red). This simulation was performed by C. Martins and E.P. Shellard

visible universe. We might then be able to detect these strings through their gravitational effects.

Strings can act as gravitational lenses - light rays propagating to us from a galaxy located behind the string will bend, resulting in two images of the same galaxy. The typical angular separation between the images is proportional to the string mass per unit length μ ; for a GUT-scale string it is a few arc seconds. The main difference from gravitational lensing by galaxies is that the two images produced by a string are expected to be nearly identical, while galactic gravitational lenses amplify and distort the images in different ways. Moving strings could also produce a characteristic signature in the CMB: the intensity of cosmic radiation would change discontinuously across the string. None of these effects have yet been observed.

Gravitational waves emitted by oscillating loops add up to a stochastic (or random) gravitational wave background with a very wide range of wavelengths—from microns to light years. The intensity of gravitational waves depends on the mass parameter μ . The fact that no gravitational

waves have yet been detected implies $\mu < 10^{19}$ kg/m. This corresponds to $E_{sb} < 10^{15}$ GeV, somewhat below the GUT scale.

14.6.3 Magnetic Monopoles

Monopoles are point-like defects. In many ways they are similar to elementary particles, but they have an unusual feature: each monopole carries a positive (north) or negative (south) magnetic charge. In contrast, all magnets we are familiar with are magnetic dipoles - they always have both a south and a north pole. The mass of a magnetic monopole is set by the symmetry breaking energy scale: $M \sim E_{sb}/c^2$. Thus, for a GUT monopole, $M \sim 10^{16}$ GeV.

As with other vacuum defects, at the time of formation, we expect to have no less than one monopole per horizon region. But unlike other defects, monopoles are predicted in *all* GUTs. This leads to a very serious problem, which we will address in Chap. 15.

14.7 Baryogenesis

At very early times, the primordial fireball was populated by particles and antiparticles in almost equal amounts. But as we noted in Sect. 14.4, there must have been a small imbalance: particles outnumbered antiparticles by about one part in a billion. What was the origin of this tiny asymmetry? Could it have been generated by some physical process from a preceding state containing exactly equal amounts of matter and antimatter?

One obstacle to implementing this idea is baryon number conservation. Protons and neutrons are collectively called baryons, and the baryon number B is defined as the number of baryons minus the number of antibaryons. A universe with equal numbers of baryons and antibaryons would have $B = 0$. To date, all observed particle processes have been found to conserve baryon number. If baryon number conservation is a universal law of nature, then B will always remain zero if it is initially zero. However, according to grand unified theories, the baryon number is only approximately conserved.¹¹

¹¹One consequence of this is that protons are not absolutely stable and can decay via processes like $p^+ \rightarrow e^+\gamma\gamma$. The expected proton lifetime is much greater than the age of the universe, but proton decay can in principle be observed by watching a huge number of protons. However, all attempts to observe it so far have failed and have led only to the upper bound of 10^{34} years on the proton lifetime.

Violations of B -conservation are extremely rare at energies that can be reached in accelerators, but they are expected to be very common in particle collisions at GUT energies. This engenders the possibility of *baryogenesis*—the generation of a nonzero baryon number in the early universe.

B non-conservation is a necessary, but not a sufficient condition for baryogenesis. If the initial state has equal numbers of particles and antiparticles, then the source of asymmetry must be in the laws of physics that govern the subsequent evolution of that state. In other words, the laws of physics should not be completely symmetric between matter and antimatter. In fact, such an asymmetry has already been observed in accelerator experiments. For example, the decay rates of the short-lived B^0 mesons are different for particles and antiparticles.

Yet another condition for baryogenesis is that B -violating reactions should be sufficiently slow, so that thermal equilibrium does not have time to establish. The reason is (roughly) that, in the absence of B -conservation, the density of each kind of particle in equilibrium is determined only by their mass. Since particles and antiparticles have the same mass, it follows that they have equal densities in equilibrium.

To see how baryogenesis may occur in the absence of equilibrium, consider the following scenario. Suppose some hypothetical X -particles and their antiparticles have asymmetric B -violating decays, so that the decay products of a particle and an antiparticle have a positive net baryon number. In the early fireball, the particle density is very high, so X -particles frequently collide and annihilate, and particle-antiparticle pairs are frequently produced in collisions of photons (and other particles), so equilibrium is established. When the temperature drops below the X -particle mass, the photon energies are no longer sufficient to produce X -pairs. Also, the particle density has now significantly decreased, so the collisions between X -particles are rare and their annihilation is inefficient. The surviving X -particles are now out of equilibrium. They eventually decay and generate a nonzero baryon number.

The three conditions for baryogenesis—baryon non-conservation, particle-antiparticle asymmetry in the laws of physics, and non-equilibrium—were first formulated in 1967 by Andrei Sakharov—the Russian physicist who is mostly known for his role in the development of the Soviet hydrogen bomb and later as a prominent dissident opposing the Soviet regime. Since then, cosmologists have suggested a number of models where these conditions are satisfied, so the observed baryon number can be generated. We don't know which, if any, of these models is correct, but most cosmologists

accept the general idea—that the matter excess over antimatter was generated by B -violating processes in the early universe.

Summary

To understand the early universe we need to understand the physics of the microworld. Around one second after the big bang, the universe was a hot fireball of electrons, protons, neutrons, photons and neutrinos. As we go farther back in time, the fireball gets hotter and denser and gets populated with other types of particles. We now have a very successful theory, called the Standard Model, which accurately describes all known particles and their interactions. There are, however, strong indications that the Standard Model is not the whole story. In particular, it does not account for some properties of neutrinos and for the existence of new (unknown) particles that constitute the dark matter.

Extensions of the Standard Model are inspired by the idea that the four fundamental forces—gravity, electromagnetism, and the strong and weak nuclear forces—are in fact different manifestations of a single, unified force. At very high energies distinctions between the different forces disappear; this is what happens at extremely high temperatures in the early universe. As the universe cools down, the symmetry between the forces is broken in several steps.

Even though the cosmic symmetry breaking transitions occurred when the universe was only a fraction of a second old, they might have left some remnants, which could be present in the universe today. These include point-like magnetic monopoles, line-like strings, and sheet-like domain walls.

Grand Unified Theories may be able to explain how an excess of matter over antimatter could have been created. GUTs predict that baryon number is not conserved, allowing for a nonzero baryon number to be generated in the early universe.

Questions

1. Why are particle accelerators so useful to cosmologists?
2. What is a positron? How are its mass and charge related to those of an electron?
3. Name two key physical properties that must be conserved when particles are produced in collisions.
4. Can two protons collide and produce two photons only? Why/Why not?

5. Can two photons of energy 0.1 MeV collide and produce an electron-positron pair?
6. List the four known fundamental forces of nature. Also indicate what role each of these forces has in nature and which bosons are responsible for mediating each force.
7. Since like charges repel, why do nuclei with many protons stay together instead of exploding apart?
8. What mass would all Standard Model particles have if the Higgs field were zero?
9. Briefly describe what happened immediately before and after the following two epochs in the very early universe:
 - (a) The electroweak phase transition ($t \sim 10^{-10}$ s, $T \sim 10^{15}$ K)
 - (b) Quark confinement ($t \sim 10^{-6}$ s, $T \sim 10^{13}$ K)
10. During the quark confinement era, protons and neutrons are formed for the first time, and most of them undergo annihilation with their anti-particles shortly thereafter. What happens to the surviving protons and neutrons? When did the protons and neutrons in your body originate?
11. Prior to the electron-positron annihilation era, there had to have been an excess of electrons over positrons. Roughly how big was this excess?
12. What is the difference between a scalar field and a vector field? Can you give an example of each?
13. In what sense is a snowflake less symmetric than a spherical drop of water?
14. Name two key features of our universe that the Standard Model does not include.
15. Which two seemingly disparate phenomena did Maxwell unify in his theory? What was his theory called?
16. Which two phenomena does the electroweak theory unify?
17. Do Grand Unified Theories include all of the forces of nature? Explain.
18. Are Grand Unified Theory (GUT) scale energies accessible to particle accelerators? If not, how can we hope to study GUT scale phenomena?
19. Some particle physics models predict the existence of defects called domain walls. What is a domain wall?
20. Why should particle physics models that predict domain walls be ruled out?
21. Do cosmic strings have ends?
22. What property of cosmic strings causes them to move with a velocity approaching that of light?
23. In string simulations, do strings become more or less wiggly with time?

24. Name one method physicists use to look for cosmic strings.
25. What is a magnetic monopole?
26. Is the statement “Diamonds are forever” consistent with grand unified theories?”

Part II

Beyond the Big Bang

15

Problems with the Big Bang

The hot big bang cosmology that we have discussed so far has been a very successful theory. It describes cosmic evolution starting from a fraction of a second after the big bang, accurately predicts the primordial nuclear abundances and the properties of the microwave background radiation, and explains how galaxies and clusters were formed over billions of years. And yet this theory fails to address some puzzling questions about our universe. Why is the geometry of the universe so close to being flat? Why is the universe so homogeneous on large scales? What is the origin of the small density fluctuations that seeded structure formation? And why is the universe expanding?

These questions do not have answers within the big bang cosmology. It simply postulates that the universe started out in a state of homogeneous expansion and was nearly flat from the start. But it is very hard to understand how such an initial state could arise, as we shall now discuss.

15.1 The Flatness Problem: Why is the Geometry of the Universe Flat?

The universe we observe today is close to having a flat, Euclidean geometry. This is equivalent to the statement that today the average energy density is nearly equal to the critical density, or that the present density parameter $\Omega_0 = \frac{\rho_0}{\rho_{c,0}}$ is close to unity. Observations indicate that Ω_0 deviates from one

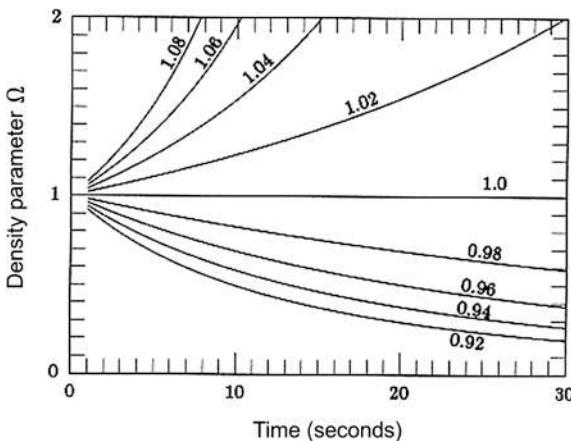


Fig. 15.1 Each line shows the evolution of the density parameter starting from the indicated initial value. In all cases, the evolution begins at one second after the big bang. If Ω was ever slightly greater than 1 in the past, then it will grow toward infinity. If Ω was ever slightly less than 1 in the past, then it declines toward zero. Thus, in order for Ω to be close to unity today it must have started out *extremely* close to unity in the past Credit Alan Guth

by no more than 1% (see Sect. 9.2). Trying to understand this major feature of our universe reveals a mystery, which is known as the flatness problem.

If the universe starts out with $\Omega = 1$, it will remain this way indefinitely. However, any slight initial deviation from unity will be amplified with time causing Ω to grow unchecked or plummet to zero¹ (see Fig. 15.1). In other words, $\Omega = 1$ is a point of unstable equilibrium. If, for example, we had $\Omega = 1.01$ at the onset of nucleosynthesis ($t = 1$ s ABB), then in less than a minute we would have $\Omega = 2$ and in a little over three minutes the universe would collapse to a big crunch. Similarly, if we started with $\Omega = 0.99$ at $t = 1$ s, then in about a year the density would be 300,000 times smaller than critical ($\Omega = 0.000003$). No galaxies or stars would ever be formed in such a low-density universe. In order for Ω to have its observed value at present, its value at $t = 1$ s must be set equal to one with an accuracy of one part in 10^{16} (see the Appendix).

Thus the flatness problem is the realization that the universe must be launched with an Ω that is finely tuned to unity, even though the big bang model cannot explain why this should be the case. It simply must be assumed as an initial condition.

¹This is only true if the universe expands with deceleration, which is indeed the case for the radiation and matter dominated epochs of the standard big bang model.

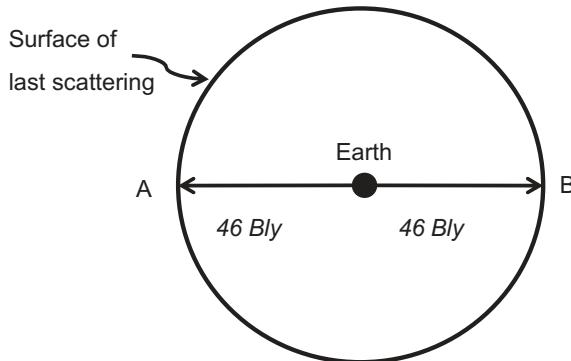


Fig. 15.2 CMB photons propagate to us from the surface of last scattering. In the standard big bang model, patches of the CMB at the points A and B have never been in causal contact. So why do they have almost identical temperatures?

15.2 The Horizon Problem: Why is the Universe so Homogeneous?

The near-uniformity of the CMB temperature over the sky tells us that the universe was extremely homogeneous at the time when the radiation was emitted. However, within the big bang model there is no reason why this ought to be the case. In fact, this ought *not* to be the case, unless the universe miraculously started out with very special initial conditions.

At first sight, a uniform temperature may not seem very surprising. A hot cup of tea left on a counter gradually cools down to room temperature. The CMB temperature could similarly equilibrate if there was some interaction between neighboring regions that emit radiation. However, when the CMB was emitted, the time that had elapsed since the big bang was too brief for such an interaction to have occurred. This is known as the *horizon problem*.

Consider the radiation coming to us from two small regions, A and B, which are diametrically opposite one another on the sky (see Fig. 15.2). The present distance to each of these regions is the distance to the surface of last scattering d_{ls} . Since the CMB radiation was emitted so early in the universe's history, there is not much difference between d_{ls} and the horizon distance d_{hor} . Thus the present distance to regions A and B is approximately equal to the horizon distance, $d_{hor} \approx 46 \text{ Bly}$. The regions are therefore separated by twice this distance and cannot possibly interact. In particular, they cannot exchange heat to equalize their temperature—and yet they are observed to have equal temperatures, up to one part in a hundred thousand.

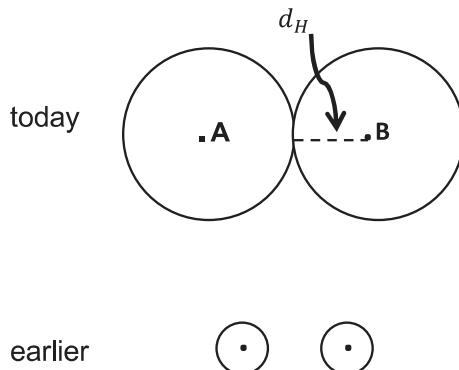


Fig. 15.3 Today regions A and B are separated by two horizon distances. Circles indicate the horizon distance from a region (black dots). As we go to earlier times, the regions get closer, but they are separated by an even larger number of horizon distances, because the horizon shrinks faster than the separation between the regions

Since the universe is expanding, regions that are distant today must have been in much closer proximity in the past. This, however, does not help to solve the problem. In fact, the horizon problem in the early universe is even more severe than it is today. To understand why, let us see how the separation between the regions d_{AB} and the horizon distance d_{hor} vary with time. Firstly, let's simplify the discussion by assuming that the universe has been in the matter dominated era² from the time the regions A and B emitted their radiation (at recombination) until the present. This means that their distance grows like the matter era scale factor $d_{AB}(t) \propto t^{\frac{2}{3}}$, and the horizon distance grows as $d_{hor}(t) \propto ct$. The horizon grows faster with time than the distance, which implies that as we go backwards from the present to the time of recombination, the horizon distance shrinks faster (see Fig. 15.3). It follows that if the separation distance d_{AB} exceeds the horizon, this excess could only be greater at earlier times. So, if two regions are now out of causal contact, they could not have been in causal contact before. For example, the regions A and B indicated in Fig. 15.2, which are now separated by $d_{AB} \approx 2d_{hor}$, were separated by approximately 80 horizon distances when the CMB radiation was emitted (see Question 5 at the end of this chapter). And at $t = 1$ s ABB they were separated by about 10^8 horizon distances.

²We disregard the recent period of accelerated expansion due to the dark energy. If we took this into account, our conclusions would remain the same.

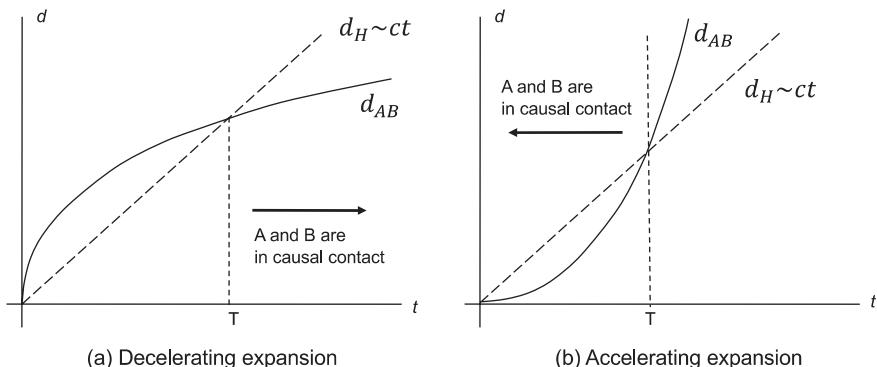


Fig. 15.4 **a** In a decelerating universe the slope of the curve $d_{AB}(t)$ decreases with time, while the horizon distance d_H has a constant slope. Regions A and B can only be in causal contact if $d_{AB} < d_H$. Initially, the universe expands so rapidly that $d_{AB} > d_H$. But, because of the decreasing slope of the curve $d_{AB}(t)$, there comes a time, denoted T , when it crosses the *straight line* representing the horizon. Once this happens, $d_{AB} < d_H$, and thus A and B are in causal contact at all later times. **b** In an accelerating universe, regions that are not in causal contact today (represented by times after T), were once in causal contact in the past (times earlier than T)

This means that hundreds of millions of causally separated regions would have had to have spontaneously started out in near perfect equilibrium in order for the CMB sky to have the uniform temperature distribution observed today.

Note that if one does assume such a special initial condition for the universe, then the hot big bang model is fully consistent. Neither the horizon nor the flatness problems are in contradiction with the big bang. They are problems in the sense that these remarkable features of the universe have no explanation within the theory.

The root of the horizon problem is that the expansion of the universe decelerates with time, so objects that are not in causal contact today could never have been in causal contact before. This makes one wonder what would happen if the universe underwent a stage of accelerated expansion. In such a universe, regions which are currently not in causal contact would in fact have been in causal contact at earlier times (see Fig. 15.4). So there would be no horizon problem. Moreover, the flatness problem would disappear as well! It can be shown that an accelerated expansion drives the value of Ω to one, even if initially it is significantly different from one. But what could cause accelerated expansion? You will have to wait until the next chapter to find out.

15.3 The Structure Problem: What is the Origin of Small Density Fluctuations?

We have marveled at the degree to which the universe is homogeneous. But even if we somehow explain the homogeneity, how do we account for the cosmic structures like galaxies, clusters of galaxies, and superclusters? In the big bang theory we have to postulate the existence of small density fluctuations, which gradually evolve into these structures. But what is the origin of the small initial density fluctuations?

15.4 The Monopole Problem: Where Are They?

As we discussed in Chap. 14, all GUT's predict that magnetic monopoles are produced during the big bang. Their initial density should be roughly one per horizon, which would result in a present density of about one monopole per cubic meter. This is comparable to the present number density of protons. But a monopole is much heavier than a proton (by a factor of 10^{16}). Thus, if monopoles were present at this density, their mass would far exceed the total mass in atomic and dark matter, in glaring conflict with observations.³ This conundrum is known as the magnetic monopole problem.

Looming behind these and other problems is an even greater mystery: what actually happened at the big bang? What was the nature of the primordial force that launched the expansion of the universe and sent particles flying away from one another? All of these questions are addressed in the theory of cosmic inflation, to which we next turn.

Summary

The hot big bang theory is supported by a wealth of observational data, but it leaves unanswered some very intriguing questions about the initial state of the universe. For example, why is the geometry of the universe today so close to being flat? The flatness problem is exacerbated by the fact that geometry tends to veer away from flatness in the course of cosmic expansion. Hence, the universe had to be extremely close to flat at very early times.

³Our density estimate here is for GUT monopoles, but the problem persists even if the monopoles are formed at a lower energy scale.

Then there is the horizon problem: observations of the cosmic background radiation indicate that the early fireball was homogeneous on scales much greater than the horizon. Since no interactions can propagate faster than light, it seems that this homogeneity could not have been established by any causal process.

Other outstanding questions include: Why was the early universe expanding? What is the origin of small inhomogeneities that later evolved into galaxies? Where did all the magnetic monopoles go? What was the universe doing before the big bang?

Questions

1. In your own words, describe what the structure, horizon, and flatness problems are.
2. Do the flatness or horizon problems contradict the big bang theory? If so, explain. If not, explain the sense in which they are “problems”.
3. Consider the graph in Fig. 15.1, showing how the density parameter Ω evolves with time. If $\Omega \approx 1$ today, what can we deduce about its value in the early universe?
4. In the hot big bang theory, does the universe undergo decelerated or accelerated expansion?
5. (a) Consider the regions A and B indicated in Fig. 15.2, which are now separated by 92 Bly. What was the distance d_{AB} between these regions at the time of recombination, $t_{rec} \approx 380,000$ yrs, when the radiation was emitted? (You may find the following facts useful: (i) the present CMB temperature is $T_0 \approx 3$ K; (ii) its temperature at recombination is $T_{rec} = 3000$ K; (iii) the temperature changes with the scale factor according to $T \propto \frac{1}{a}$.)
(b) Find the ratio d_{AB}/d_{hor} at the time of recombination, $t_{rec} \approx 380,000$ yrs. You may use the equation $d_{hor} \sim 3ct$ (introduced in Chap. 7) to approximate the horizon distance at recombination.

16

The Theory of Cosmic Inflation

The horizon and flatness problems had been recognized since the 1960s, but were rarely discussed—simply because no one had any idea as to what to do about them. These problems could not be resolved without addressing the question of what really happened at the earliest moments of the big bang. With no progress in that direction, physicists grew accustomed to the notion that questions about the initial state of the universe belonged to philosophy, not physics. It therefore came as a total surprise when, in 1980, Alan Guth made his dramatic breakthrough, providing a way to resolve the stubborn cosmological puzzles in one shot.

16.1 Solving the Flatness and Horizon Problems

The origin of the flatness and horizon problems can be traced to the *decelerated expansion* of the universe. In a decelerating universe the density parameter Ω is driven away from one, and thus it is remarkable that it is measured to be so close to unity today. Also, if the expansion decelerates, the horizon grows faster than the separation between regions. This means that if we look backwards in time, the horizon shrinks faster than the separation between any two regions. Thus regions that are not in causal contact now, could never have been in causal contact at any earlier time. Both of these problems can be solved if the universe underwent a period of *accelerated expansion* in its infancy. But *what could have caused such a period?*

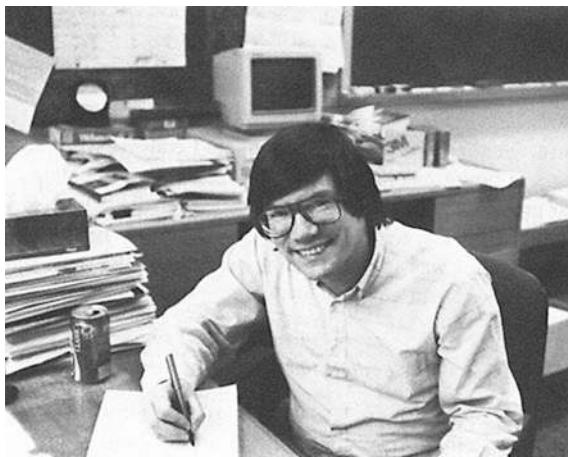


Fig. 16.1 Alan Guth came up with his idea of cosmic inflation while he was in his 9'th year as a temporarily employed postdoctoral fellow. Soon thereafter, he became a tenured Professor at MIT. For his work on inflation Guth won the 2012 Fundamental Physics Prize. Among his other awards, he also won the 2005 contest for the messiest office in the Boston area (organized by the local newspaper *The Boston Globe*)

You might have guessed the answer—vacuum energy! We know that the expansion of the universe is now accelerating, due to the repulsive gravity of the vacuum. However, this accelerated expansion only began at a relatively recent cosmic time, when the density of matter ρ_m dropped below the vacuum energy density ρ_v . At earlier epochs ρ_v was totally negligible so it could not have caused accelerated expansion in the early universe. What we need is a vacuum with a huge energy density at very early times. Fortunately, grand unified theories of particle physics make the existence of such high-energy vacuum states plausible. This led Alan Guth to propose that a large vacuum energy density caused the universe to undergo a period of very fast, accelerated expansion, thereby solving the flatness and horizon problems. Guth also suggested a fitting name for the accelerated expansion epoch: *cosmic inflation* (Fig. 16.1).

16.2 Cosmic Inflation

16.2.1 The False Vacuum

Vacuum is what you end up with when you remove all that can be removed. It is empty space. But according to modern particle physics, vacuum is very

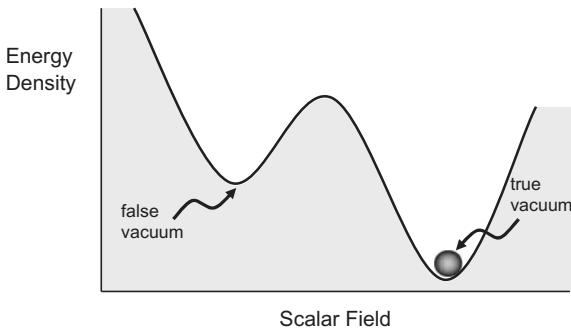


Fig. 16.2 Potential energy density of a scalar field. Here there are two minima, one of which is the true vacuum

different from nothing. At every point in space there is a Higgs field, as well as other scalar fields, responsible for the grand unified symmetry breaking. The vacuum values of these fields determine the masses and interactions of all elementary particles. Symmetry can generally be broken in several different ways, and thus we expect to have a number of vacuum states with different properties. Particle physicists refer to these states as different *vacua*.

As an illustration, let us consider a toy “grand unified theory” with a single scalar field that has a potential energy density curve shown in Fig. 16.2. We can represent the value of the field by a ball that rolls in this energy landscape and comes to rest in one of the two valleys. The valleys represent the two possible vacuum states in this theory. The lowest-energy vacuum is the absolute minimum of energy; it is called the “true vacuum”. Any higher-energy vacuum is necessarily unstable; hence it is called a “false vacuum”. We know that physical systems tend to minimize their potential energy, so a false vacuum has to decay by converting into true vacuum. (We shall discuss the decay process later in this chapter.)

We are now ready to formulate the idea of cosmic inflation, as it was originally proposed by Guth. Suppose the universe was in a high-energy false vacuum state at some early time in its history. The strong repulsive gravity of the false vacuum would then cause a period of very fast, accelerated expansion. This would solve the horizon and flatness problems of the standard big bang cosmology. The inflationary period ends when the false vacuum decays into the true vacuum. The excess energy of the false vacuum has to go somewhere, and Guth assumed that it gets converted into a hot fireball of particles. The fireball continues to expand by inertia, and the expansion rate gradually slows down due to gravity. The end of inflation plays the role of the big bang in this scenario. At later times, the universe evolves along the lines of the hot big bang cosmology.

How large can we expect the false vacuum energy density to be?

The answer depends on the details of particle physics, but we can make an educated guess, using a trick called "dimensional analysis". Each particle physics model has a characteristic mass, or energy scale, which we shall denote by M . For the electroweak theory, this scale is $M \sim 100 \text{ GeV}$. The Higgs, W and Z boson masses all have this order of magnitude. The electroweak symmetry breaking energy has a similar magnitude: electromagnetic and weak interactions cannot be distinguished at particle energies much greater than 100 GeV . For grand unified theories the corresponding mass/energy is $M \sim 10^{16} \text{ GeV}$. This mass also determines the characteristic scale of the energy density landscape of the theory, that is, the typical height of the hills and valleys in Fig. 16.2. We can thus expect to have a formula expressing the false vacuum energy density ρ_v in terms of M and of the fundamental physics constants—the Planck constant \hbar and the speed of light c . Now, the key point is that only one combination of M , \hbar and c has the dimension of energy density; it is

$$\rho_v \sim \frac{c^5 M^4}{\hbar^3} \quad (16.1)$$

There may also be a numerical coefficient, but it usually does not change the order of magnitude by too much. The above formula can be rewritten as

$$\rho_v \sim 10^{21} M_{\text{GeV}}^4 \frac{\text{kg}}{\text{m}^3}, \quad (16.2)$$

where M_{GeV} is the mass M expressed in units of GeV. For a grand unified theory with $M_{\text{GeV}} \sim 10^{16}$, this gives a truly enormous density of $\rho_v \sim 10^{85} \text{ kg/m}^3$. One cubic centimeter of this vacuum contains much more energy than our entire observable universe!

16.2.2 Exponential Expansion

While the universe stays in the false vacuum state, the energy density remains constant. This leads to a very special kind of growth, called exponential expansion. The hallmark of exponential expansion is that in a fixed period of time t_D (the doubling time) the size of a given region will double (and its volume will therefore be increased by a factor of $2^3 = 8$). So, if we start with one cubic nugget of vacuum with length l_0 , after one doubling time the cube will have size $2l_0$. After the next doubling time it will have size $2 \times 2 \times l_0 = 4l_0$; and after n doubling times it will have size $2^n l_0$. This is similar to financial inflation at a constant rate. A remarkable property of exponential growth is that the numbers get enormous after relatively few doubling cycles. For example, if a slice of pizza costs \$1 now, then after n doubling cycles it will cost $\$2^n$: so after 10 cycles it will cost \$1024, and after 330 cycles it will cost $\$10^{100}$ (this much money does not even exist) (Fig. 16.3).

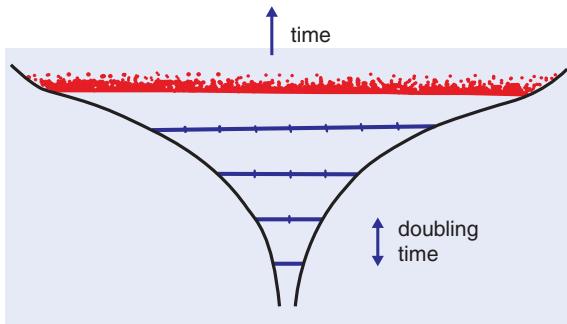


Fig. 16.3 Exponential expansion of the universe. At each time step the universe doubles in size

A universe undergoing exponential expansion is also characterized by a Hubble parameter which does not change in time, $H = \text{const.}$ This is not difficult to understand, if we recall that an accelerated expansion drives the density of the universe towards the critical value,

$$\rho_c = \frac{3H^2}{8\pi G}. \quad (16.3)$$

During inflation we have $\rho = \rho_v$; then, setting $\rho_c = \rho_v$ and solving for H , we obtain

$$H = \left(\frac{8\pi G \rho_v}{3} \right)^{1/2} = \text{const.} \quad (16.4)$$

Any two particles in the inflating universe are driven apart with a velocity given by Hubble's law, $v(t) = Hd(t)$, where $d(t)$ is the distance between the particles. Let's say at some moment they are separated by distance d and receding with speed v (at any later time both the distance and velocity increase). If the particles were to continue separating at this speed, then the distance between them would double in a time interval $t_D = d/v = 1/H$. Since the universe expands with acceleration, the actual doubling time is somewhat shorter, but the relation $t_D \sim 1/H$ still gives a good order of magnitude estimate.¹

If inflation happened at the GUT-scale, then $H \sim 10^{38} \text{ s}^{-1}$ and $t_D \sim 10^{-38} \text{ s}$. With such an incredibly brief doubling time, the universe would expand by a factor of 10^{100} in less than 10^{-35} s . If, for example, we start with

¹We show in the Appendix that the doubling time in an inflating spacetime is $t_D = \frac{0.7}{H}$.

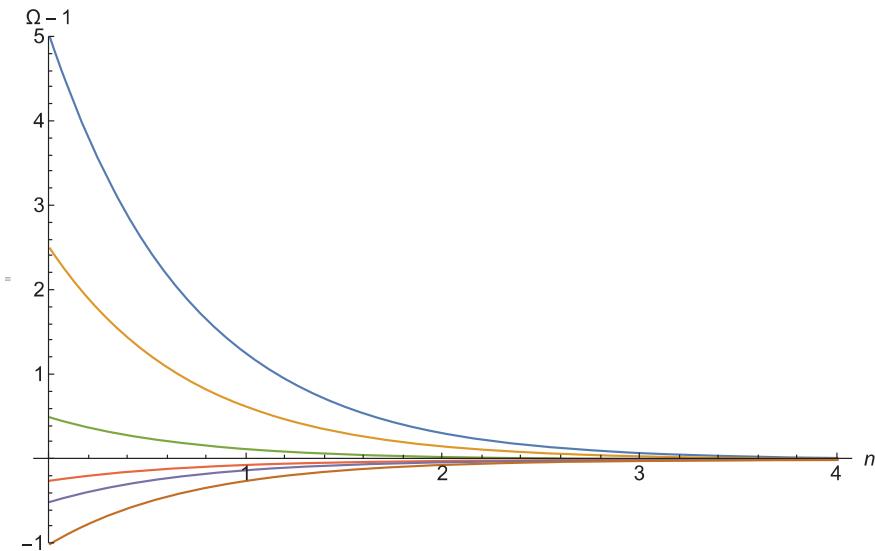


Fig. 16.4 Evolution of the density parameter in an inflationary universe. Here we plot $\Omega - 1 \propto 2^{-2n}$, where n is the number of doubling times (see the end of the appendix for derivation of this equation). Above the x-axis the curves are for closed universes, with different initial values of Ω ; below the x-axis the curves are for open universes with different initial values for Ω . Even if the density parameter starts off much larger or smaller than unity, it is quickly driven to unity within several doubling times of inflation

a nugget of false vacuum having the size of a proton, $\sim 10^{-15}$ m, then in less than 10^{-35} s the universe would inflate to a size of 10^{85} m. This is vastly larger than the size of the observable universe, which is “only” 10^{26} m. The exponential expansion of the false vacuum is thus an immensely powerful mechanism that can blow a tiny seed universe up to astronomical dimensions in a very short time.

16.3 Solving the Problems of the Big Bang

Let us now see how inflation helps to explain the puzzling features of the initial state that had to be postulated in the big bang theory.

16.3.1 The Flatness Problem

A period of exponential expansion drives the density parameter towards $\Omega = 1$. Figure 16.4 illustrates that Ω gets extremely close to one in a relatively small number of doubling times. This means that the geometry of the universe gets very close to flat, Euclidean geometry.

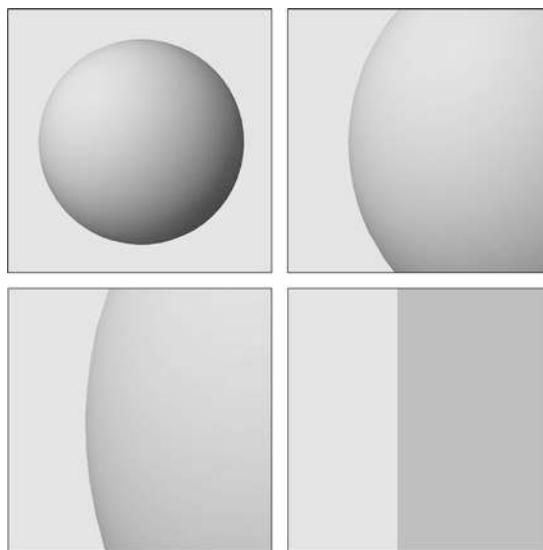


Fig. 16.5 The surface of a huge balloon looks flat, because we can see only a small part of it. Similarly, the universe appears to be flat because we only see a small portion of it after inflation

This effect has a simple intuitive explanation. Imagine a curved surface, like a sphere. Now imagine enlarging this surface by a huge factor. This is what happens to the universe during inflation. We can now see only a tiny part of this big universe. And it appears to be flat, just like the surface of the Earth looks flat when we see a small portion of it (Fig. 16.5).

16.3.2 The Horizon Problem

Consider a spherical region of diameter d much smaller than the Hubble distance d_H at the beginning of inflation. The region is initially expanding at a much smaller speed than light, so there is plenty of time for different parts of the sphere to interact and come to equilibrium. Then the inflationary expansion blows the size of the region up by a factor 2^n , where n is the number of doubling times during inflation. This factor can be enormous, so the present size of the region can easily be much larger than our observable universe. This solves the horizon problem: the CMB temperature is uniform

over the sky because all parts of the observable universe were in causal contact at the beginning of inflation.

What is the minimal number n_{\min} of doubling times that is necessary to solve the horizon and flatness problems? The answer depends on the energy scale of inflation, M . For M at the GUT scale, one finds n_{\min} is about 90.

16.3.3 The Structure Formation Problem

Inflation also offers the most promising explanation for the origin of small density fluctuations that later evolved into galaxies and clusters. We will discuss this shortly.

16.3.4 The Monopole Problem

All monopoles produced before or during inflation get diluted away by the huge inflationary expansion, so that their present density becomes negligible.

16.3.5 The Expansion and High Temperature of the Universe

The hot big bang model assumes that the universe started out in a state of rapid expansion at a very high temperature. But why was the early universe so hot? And why was it expanding? Inflation provides a possible explanation for this initial state. The expansion of the universe is caused by the repulsive gravity of the false vacuum. The vacuum energy density during inflation is expected to be very high, and when the false vacuum decays, this energy gets converted into a hot fireball of particles and radiation; hence the fireball is born with a very high temperature.

We thus see that a period of inflation in the early universe can resolve the perplexing problems of the big bang. But in order for the inflationary model to be complete, we need to understand how inflation begins and how it ends. Inflation ends when the false vacuum decays, so in the next section we shall study the vacuum decay process. The question of the beginning of inflation will be addressed in Chap. 23.

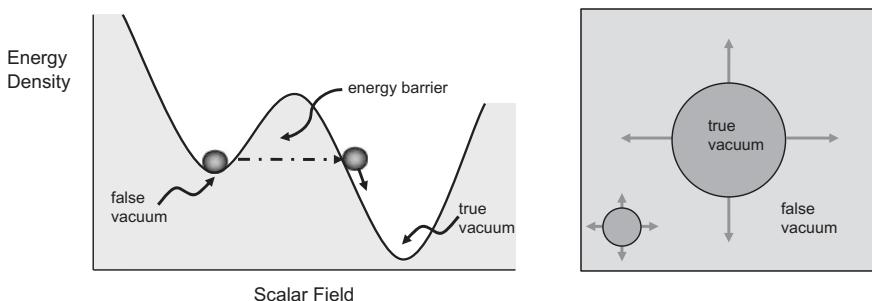


Fig. 16.6 Vacuum decay. When the field tunnels from its false vacuum to its true vacuum value, bubbles of true vacuum nucleate within the false vacuum background. The bubbles then expand at speeds approaching the speed of light

16.4 Vacuum Decay

16.4.1 Boiling of the Vacuum

Consider the energy landscape of a scalar field illustrated in Fig. 16.6. It has a false vacuum and a true vacuum. During inflation the field is in the false vacuum everywhere in space. Now, in order for the false vacuum to decay, the field has to overcome the energy barrier separating the two vacua. As we discussed earlier, the dynamics of the scalar field is similar to that of a ball rolling in the energy landscape. If the ball is located in the valley marked “false vacuum”, then, according to classical physics, it will stay there forever, unless someone kicks it upwards, providing the energy needed to go over the barrier. But we learned in Chap. 10 that the ball can quantum-mechanically tunnel through the barrier and emerge on the other side. This is also what happens in vacuum decay.

Quantum tunneling is a probabilistic process, so you cannot predict exactly when and where it is going to happen. You can only calculate the probability for tunneling to occur in a given region of space per interval of time. The probability for a large region of false vacuum to tunnel to the true vacuum is extremely low. Thus tunneling occurs in a tiny microscopic region, resulting in a small true vacuum bubble.²

The process of vacuum decay is similar to the boiling of water. Small bubbles of true vacuum pop out (or “nucleate”) randomly in the midst of false

²Despite the similarity between the tunneling of a ball and that of a scalar field, there is also an important difference. The ball tunnels between two different points in space, while the field tunnels between two different field values at the same location in space.

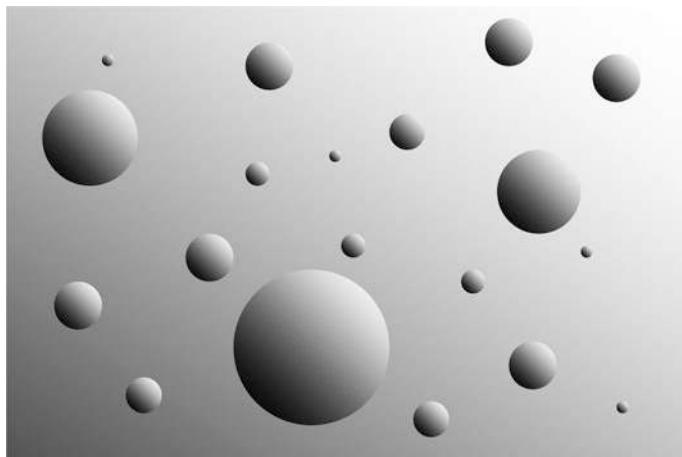


Fig. 16.7 Inflating universe with true vacuum bubbles. Bubbles are driven apart by the expansion of the universe, so they almost never collide

vacuum. The energy released by converting false vacuum into true vacuum gets concentrated in the bubble walls, which expand at a speed approaching the speed of light. When bubbles collide and merge, the walls disintegrate into particles. This is how Guth originally envisioned the end of inflation and the onset of the big bang. But unfortunately this broad-brush scenario has a fatal flaw.

16.4.2 Graceful Exit Problem

The problem is that even though the bubbles expand at nearly the speed of light, we cannot simply assume they will collide, because the space between them is filled with false vacuum and is also rapidly expanding. In fact, any bubbles separated by more than a Hubble distance d_H are driven apart faster than the speed of light and will never collide. The typical distance between the bubbles depends on the rate at which they nucleate. If the nucleation rate is low, then bubbles will form separated by wide stretches of false vacuum and will almost never collide. All the energy of the bubbles will remain concentrated in the expanding bubble walls, and inflation will never end (Fig. 16.7).

To get around this problem, we can consider a model where bubbles nucleate at a very high rate, so that their typical separation is less than d_H . In this case the bubbles will collide and merge, and the whole vacuum decay process will be over in less than a doubling time. But in order to solve the

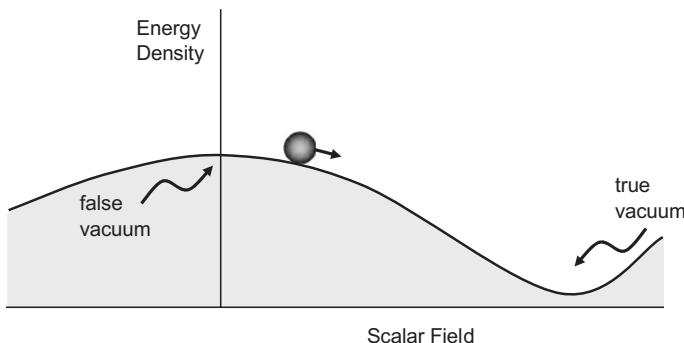


Fig. 16.8 In the slow roll inflation scenario, the role of the false vacuum is played by a very flat plateau at the top of the potential energy density hill

horizon and flatness problems we need inflation to persist for many doubling times (roughly 90 or so, depending on the details of the model). Thus we are faced with an impasse: either inflation does not end at all, or it ends too rapidly to solve the problems it was invented to solve. In the early 80's this became known as *the graceful exit problem*. Guth realized that his theory suffered from this problem soon after he came up with the idea of inflation, so he concluded his landmark paper stating: "I am publishing this paper in the hope that it will encourage others to find some way to avoid the undesirable features of the inflationary scenario."

16.4.3 Slow Roll Inflation

The Russian born cosmologist Andrei Linde was the first to find a solution to the graceful exit problem in 1982. A few months later the same idea was independently proposed by Andreas Albrecht and Paul Steinhardt in the USA. The crucial step was to consider an energy landscape without a barrier, but with a very gentle slope, as shown in Fig. 16.8. Once again, we can represent the scalar field by a ball rolling in this landscape. If we place the ball near the top of the hill, it will start slowly rolling down, and since the slope is so flat, the ball will initially stay at about the same height. For the scalar field this means that its energy density will remain almost constant. But a constant energy density is all that is needed to sustain a constant rate of inflation.

The flat region near the hilltop can be called a "false vacuum". Since the field "rolls" very slowly, it takes a while for it to cross that region, and in the meantime the universe expands exponentially. Once the field reaches the



Fig. 16.9 Andre Linde has been one of the chief architects of inflationary cosmology for over 30 years. Linde began his career in his native Moscow, and has been a Professor at Stanford University since 1989. He often collaborates with his wife Renata Kallosh, who is also a Professor at Stanford. Linde is a flamboyant, entertaining presenter and an outspoken champion of his ideas. He is an excellent artist and occasionally illustrates his lectures with beautifully drawn cartoons. His numerous hobbies include swimming, juggling, card tricks, and photography. *Credit Vadim Shultz*

true vacuum, it oscillates back and forth and eventually comes to rest, with its energy turned into a hot fireball of particles.³ By this time the universe has expanded by a huge factor.

Note that in this model, the field “rolls” simultaneously at all points in space and produces a fireball at the same time in the entire inflating region. We thus have a large, hot, homogeneous, expanding universe. The graceful exit problem has been solved! (Fig. 16.9).

Like the Higgs field of the Standard Model, the rolling scalar field must have some particle associated with it. Particle physicists have suggested a number of candidates, but none of them is particularly compelling. For now, the particle goes by the generic name “inflaton”, and the field is called the “inflaton field”.

³A ball rolling on a similarly curved surface would also oscillate about the lowest point, would gradually slow down due to friction, and would come to rest, with all its mechanical energy turned into heat. Similarly, analysis shows that an oscillating field loses its energy by particle production, creating a fireball.

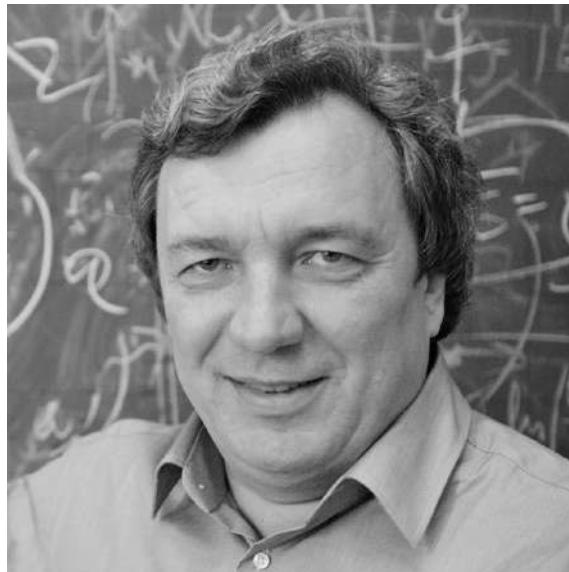


Fig. 16.10 Viatcheslav Mukhanov suggested (with Chibisov) that cosmological density fluctuations could have a quantum origin and later did seminal work developing the details of this scenario. He is known among cosmologists for his flamboyant personality and politically incorrect sense of humor. Credit PR Image iau1304a, Viatcheslav Mukhanov, recipient of the 2013 Gruber Prize (<https://www.iau.org/news/pressreleases/detail/iau1304/>)

16.5 Origin of Small Density Fluctuations

As we discussed in Chap. 12, galaxies and galaxy clusters arise from gravitational collapse in a universe that begins with small density variations from one location to the next. But where do these initial density variations come from?

If inflation gives us a perfectly homogeneous universe, then it works *too well*. Russian physicists Viatcheslav Mukhanov and Gennady Chibisov proposed in 1981 that density fluctuations in the early universe could arise due to random quantum fluctuations. This means that quantum effects, which are usually only important in the microworld, could be ultimately responsible for the existence of the largest structures in the universe! (Fig. 16.10).

Let's see how this is possible. In addition to classical motion, the inflaton field is subject to quantum mechanical effects (see Fig. 16.11). As the field rolls downhill, it experiences quantum fluctuations, which randomly kick the field *up or down* the hill. The directions of these small kicks are not the same in different spatial regions of the universe. Thus the field arrives at the bottom of the hill and produces a fireball at different times in different spatial locations.

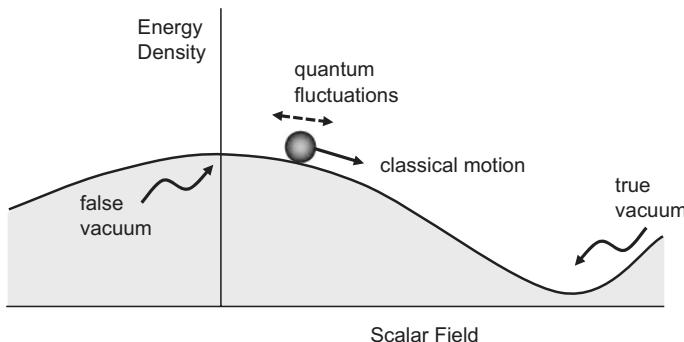


Fig. 16.11 The evolution of the inflaton field is a combination of its deterministic classical motion down the hill and its random quantum mechanical jumps up and down the hill

In regions where the field took a little longer to reach the true vacuum, inflation lasts a bit longer, and the matter density will be slightly higher. Why? Because during inflation, the energy density stays roughly constant even as the universe expands. But once the fireball is produced, matter and radiation get diluted. So parts of the universe that exit the inflating stage sooner, get a bit diluted by the time nearby lagging regions start their hot big bang evolution. The upshot of inflation ending at slightly different times is that the very early universe is imbued with small differences in density from one region to another. These are the density fluctuations that could be responsible for the formation of cosmic structure (Fig. 16.12).

All density fluctuations originate as quantum kicks in tiny regions which have a size roughly given by the Hubble distance d_H .⁴ But then they are stretched by the expansion to a much greater size. Fluctuations produced earlier are stretched for a longer time and encompass a larger region. The magnitude of fluctuations is set by the initial quantum kick and is about the same for all distance scales. This leads to a scale-invariant spectrum of density fluctuations.

To clarify what a scale invariant spectrum means, imagine that we divide the universe into cubic regions of size 100 light years and measure the aver-

⁴Quantum fluctuations occur on smaller scales as well, but upward and downward kicks alternate in rapid succession, so their overall effect is nil. But once the fluctuation region is stretched to a size larger than d_H , its different parts become causally disconnected, and coherent fluctuations in such a region are no longer possible. The surviving fluctuations are the ones produced in regions of size $\sim d_H$. The region is then immediately stretched to a larger size, and the fluctuation “freezes”.

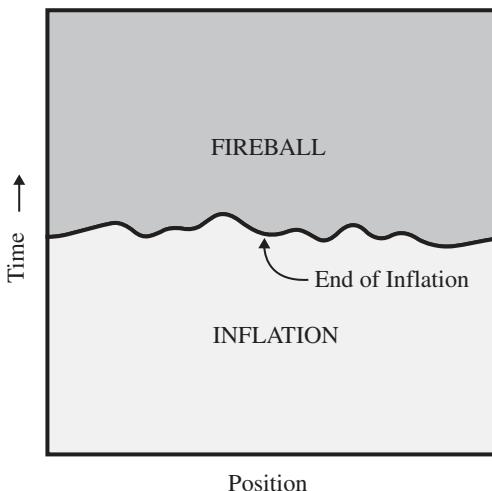


Fig. 16.12 Different ending times for inflation result in small density fluctuations. Regions where inflation ends later have a higher matter density

age density in each cube. Let's say we find the density fluctuation (that is, the typical variation from one cube to another) to be 1%. Now we can repeat this experiment with cubic regions of different size (say, 1000 light years, 10 light years, etc.). If the density fluctuation is the same for any choice of the size, we say that the spectrum of density fluctuations is scale invariant.

The scale invariance of density fluctuations from inflation is only approximate. The magnitude of quantum kicks decreases slightly as the field rolls downhill. As a result, the density fluctuations on greater distance scales, which were produced when the field was at a higher altitude, are slightly larger than the smaller-scale fluctuations. The form of the spectrum of primordial density fluctuations is one of the most important observational predictions of inflation, as we will discuss in Chap. 17.

16.6 More About Inflation

16.6.1 Communication in the Inflating Universe

Let us imagine two comoving observers in an inflating universe, who communicate by exchanging light signals. The observers are moving apart at the

speed $v = Hd$, where d is the distance between them. Suppose the observers begin their signal exchange when d is very small, so that the separation speed v is small compared to the speed of light. Then the Hubble expansion has almost no effect on light propagation between the observers, and they can exchange many signals before their distance is appreciably enlarged. But as the observers move apart, their separation speed gradually increases and becomes equal to the speed of light at the Hubble distance

$$d_H = \frac{c}{H} \quad (16.5)$$

When d gets close to d_H , light signals take longer and longer to propagate and arrive more and more redshifted. Once d becomes greater than d_H , any communication between the observers becomes impossible, since they are now moving apart faster than the speed of light. At later times v will get only larger, so light signals sent by one observer will never catch up with the other one. We thus see that comoving observers who are in causal contact in an inflating universe will necessarily fall out of causal contact at later times (assuming that inflation continues).

A spherical surface of radius d_H surrounding an observer is called the observer's Hubble sphere; its properties in an exponentially expanding universe are similar to those of the Schwarzschild horizon of a black hole. Events that occur beyond the Hubble sphere cannot be detected by the observer. For GUT-scale inflation, the Hubble distance is tiny, $d_H \sim 10^{-30}$ m, hardly enough to contain an observer. But note that today our universe is vacuum-dominated and is once again undergoing a stage of exponential expansion. The present vacuum energy density is much lower than it was in the early inflationary phase, and the Hubble distance is now astronomically large: $d_H \sim 10^{10}$ ly. Galaxies in the observable universe are driven towards d_H . As they approach the Hubble sphere, they become more and more redshifted and gradually fade away.

16.6.2 Energy Conservation

A short period of inflation can blow a tiny subatomic-size region up to dimensions much greater than the entire observable universe. On the face of it, this seems to be in conflict with energy conservation. The false vacuum has a constant energy density ρ_v , so its energy is proportional to the volume V that it occupies, $E_v = \rho_v V$. At the end of inflation the volume is enor-

mous, and so is the energy. The question is: where did all this energy come from?

To see what is going on here, let us first note that the total energy must include the contribution of the gravitational potential energy. Furthermore, let us recall that the gravitational energy is always negative and that it also gets large when the mass is large. Hence it is conceivable that as the mass/energy of the false vacuum grows during inflation, its negative gravitational energy grows at the same rate, so the total energy remains constant.

An analogous situation arises in Newtonian theory when a small particle falls towards a massive star of mass M . The energy of the particle in this case is

$$E = \frac{1}{2}mv^2 - \frac{GMm}{r}, \quad (16.6)$$

where m is the particle's mass, v is its velocity, and r is its distance from the center of the star. The first term is the particle's kinetic energy and the second is the gravitational potential energy. Suppose the particle is initially at rest ($v = 0$) at a large distance from the star, so the energy is very small. As it falls, the particle accelerates, and its kinetic energy can get very large as it approaches the star. On the other hand, as r gets smaller, the potential energy gets large and negative. But the two contributions nearly cancel one another, so the total energy is conserved and is close to zero, as it was from the start.

A detailed analysis, based on general relativity, shows that the energetics of cosmic inflation is very similar. The total energy of the huge false vacuum region at the end of inflation is very tiny; it is the same as the energy of the initial nugget from which this volume originated.

Summary

The horizon and flatness problems can be solved if the universe underwent a period of *accelerated* expansion, called inflation. The theory of inflation assumes that the universe originated in a state of a high-energy false vacuum. The repulsive gravity of that vacuum causes a super-fast, exponential expansion of the universe. Regardless of its initial size, the universe very quickly becomes huge. The false vacuum eventually decays, producing a hot fireball, marking the end of inflation. The fireball continues to expand by inertia and evolves along the lines of hot big bang cosmology. Decay of the false vacuum plays the role of the big bang in this scenario.

The theory of inflation explains the expansion of the universe (it is due to the repulsive gravity of the false vacuum), its high temperature (due to the high energy density of the false vacuum), and its observed homogeneity (false vacuum has an almost constant energy density). It also predicts a nearly scale invariant spectrum of density fluctuations, which can serve as seeds for structure formation.

Questions

- What is cosmic inflation? How is it different from a rapid expansion of the early universe in the hot big bang cosmology?
- What properties of the false vacuum are responsible for the inflationary expansion?
- If inflation occurred at the electroweak energy scale, the doubling time would be $t_D \sim 10^{-10}$ s. Roughly how long would it then take for the universe to grow by a factor of 1000?
- How does inflation solve the horizon and flatness problems?
- Does the theory of inflation replace the big bang theory? Explain.
- Name the key features of the universe that a brief period of inflation explains.
- In the context of Guth's original inflationary model: (a) In what sense is vacuum decay like the boiling of water? (b) When false vacuum is converted into a true vacuum bubble, energy is released. Where is this energy stored?
- The original version of inflation had the so-called "graceful exit problem". Even though bubbles of true vacuum expand with almost the speed of light, it was difficult to get the bubbles to collide and release their energy. What prevented the bubbles from colliding?
- Suppose we have a potential energy landscape with a steep slope and no barrier. If the scalar field starts somewhere on the slope, it will roll down very quickly. Would this give a satisfactory inflationary scenario?
- Consider the potential energy density diagram in Fig. 16.8. What key feature of this potential solved the "graceful exit problem". Briefly explain how this potential gives rise to a large, hot, homogeneous, expanding universe.
- How does inflation explain the origin of small density fluctuations?
- As the inflaton field rolls down the energy hill, it experiences random quantum fluctuations in different directions. Thus the field arrives at the bottom at different times, in different regions. For those regions where inflation lasted a little longer, will the matter density be slightly higher or lower than average? Why?

13. Quantum fluctuations take place in tiny regions of space. They are then stretched to macroscopic sizes by the expansion of space. Early fluctuations undergo more stretching and thus encompass larger regions than those produced later on. Is the magnitude of the resulting density fluctuations on large scales less than, greater than or the same as the magnitude of fluctuations on much smaller scales?

17

Testing Inflation: Predictions and Observations

We have seen how cosmic inflation can create an enormous universe from a tiny seed, while solving many problems that plagued pre-inflation models. Most cosmologists have embraced the inflationary scenario, but how can we know that inflation actually happened? Fortunately, the theory of inflation has made several testable predictions, three of which we now discuss.

17.1 Flatness

As we learned in the previous chapter, accelerated expansion during inflation rapidly drives the density parameter towards $\Omega = 1$ and the geometry of the universe to flatness. Thus, inflation predicts that, on the largest observable scales, the universe should be accurately described by a flat geometry with $\Omega = 1$. When Alan Guth made this prediction in the early 80's, astronomers viewed it with a high degree of skepticism. All the evidence at that time pointed to an open hyperbolic universe. Even including dark matter, observation favored $\Omega_m \sim 0.3$.

Then, quite unexpectedly, nearly 20 years after Guth's prediction, dark energy was discovered. Today we know from CMB and supernovae measurements that its contribution to the cosmic energy balance is $\Omega_{vac} \approx 0.69$, so that

$$\Omega_{tot} = \Omega_m + \Omega_{vac} = 1 \pm 0.01, \quad (17.1)$$

The universe is thus very close to being flat, in excellent agreement with the inflationary prediction.

17.2 Density Fluctuations

Perhaps the most impressive triumph of inflation has been the explanation of primordial density fluctuations. The theory of inflation specifically predicts that the magnitude of fluctuations is about the same for all observable distance scales: the initial density fluctuations have a scale-invariant spectrum. Also, the different components of the early universe—dark matter, electrons, protons and photons—all start out with the same density perturbations. Thus, regions that have an initial over-density of photons, say, by 0.01%, also have a 0.01% initial over-density of dark matter, and so on. (However, as we shall soon see, the dark matter does not evolve in tandem with the other components during times of interest in this chapter.)

To test this prediction, cosmologists used a computer code to follow the evolution of scale-invariant fluctuations up until recombination. The resulting pattern of evolved density fluctuations was then converted into a pattern of temperature anisotropies, and these were compared to the CMB observations. The agreement between theory and experiment, as most accurately measured by the Planck satellite (see Fig. 17.1), is striking.¹

Let us now discuss the CMB temperature anisotropies depicted in Fig. 17.1 in more detail. The horizon distance at recombination corresponds to about 1.5° on the sky (see Question 9). Matter could not have moved over distances exceeding the horizon, and thus on angular scales significantly larger than 1.5° the temperature anisotropies represent the density fluctuations in their pristine form, as they came out of the inflationary epoch. As expected, the magnitude of the anisotropies, and hence the density fluctuations, is about the same for all angles in this range.

Photons that propagate to us from higher density regions start out a little hotter than average, and those that travel towards us from less dense regions are initially cooler than average. On the other hand, photons from denser regions lose more energy as they climb out of stronger gravitational fields produced by those regions. This gravitational redshift turns out to be the dominant effect, so the net result is that, surprisingly, hot patches in

¹Recall, the scale invariance of density fluctuations from inflation is only approximate: fluctuations on greater distance scales are slightly larger than the smaller-scale fluctuations. The experimental data are consistent with these details.

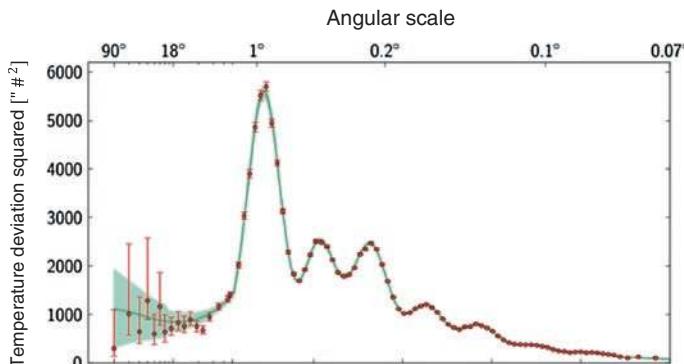


Fig. 17.1 Planck satellite temperature anisotropies. The squares of temperature deviations are plotted (in micro Kelvins squared) versus the angular scale on the sky subtended by hot or cold spots. The red dots are data points, and the green line is the theoretical prediction. On large angular scales, there are only a few cold or hot patches that can fit in the sky. This results in a large statistical uncertainty, indicated by the light green band in the figure. Credit ESA and the Planck Collaboration

the CMB sky indicate under-dense, cooler regions in the early universe at recombination.

On angular scales of about 1° and smaller, the data in Fig. 17.1 exhibit a number of peaks. These peaks are signatures of primordial sound waves, as we shall now explain. Prior to recombination, protons, electrons and photons are tightly coupled together by electromagnetic forces and act as a single proton-electron-photon gas. Such a mixture of charged particles and radiation is called a *plasma*. In denser regions the plasma has a higher temperature and a higher pressure. The difference in pressure pushes the plasma into neighboring low-density regions. The pressure momentarily equalizes, but the plasma keeps moving by inertia, so the regions that were initially hotter and denser become cooler and vice versa. Now the pressure difference works in the opposite direction and the plasma rushes back to the initially over-dense regions. The resulting oscillations of compression and rarefaction are simply sound waves in the primordial plasma. Dark matter particles interact very weakly with ordinary matter, so they do not participate in plasma oscillations.

Like sound waves in the air, the plasma sound waves are characterized by an oscillation period P and a wavelength λ , equal to the distance traveled by sound in one period. The wavelength of the plasma waves is determined by the size of the over-dense and under-dense regions. Since these regions come in a variety of sizes, many different waves are “sounding” at the same time. The speed of sound v_s in a gas is comparable to the typical particle velocity.

The cosmic plasma consists predominantly of photons, so v_s is not much different from the speed of light: $v_s \approx 0.6c$. When charged particles eventually recombine to form neutral atoms, the oscillations stop, and photons stream freely through the universe, bringing us the pattern of primordial sound, frozen in time at recombination.

The waves that reach their maximum amplitude at t_{rec} contribute the most to the observed temperature fluctuations. The longest waves of this kind have period $2t_{rec}$. Their wavelength λ_f is called the fundamental wavelength. A simple calculation shows that $\lambda_f \approx 1.4 \times 10^6$ lyrs (see Question 10). By the time of recombination, such a wave completes half a period of oscillation, so a region that started out at maximal rarefaction after inflation will have reached maximal compression at t_{rec} , and vice versa. The highest peak in Fig. 17.1 is generated by sound waves with the fundamental wavelength. Other peaks come from sound waves having wavelengths that are integer fractions of λ_f . The second peak is due to waves with half the fundamental wavelength. In this case, a maximally rarefied region has had time to reach maximum compression and then rebound, again becoming maximally rarified by the time of recombination. The third peak is due to a sound wave with a third of the fundamental wavelength, and so on. It is also evident in the figure that peaks at smaller angular scales drop off in magnitude. This is due to the dissipation of sound waves.

Mining the CMB

There is much that can be learned about our universe from the CMB anisotropies. The angle of the fundamental peak in Fig. 17.1 allows cosmologists to directly measure the curvature of the universe. This angle (approximately 1°) gives us the angular size of the most intense temperature variations on the sky.

It is the angle subtended by half of the fundamental wavelength λ_f (because the full wavelength includes both a cold and a hot patch). Since we know both the distance from us to the surface of last scattering (this distance is very close to the horizon distance; see Sect. 7.7.) and the physical size of λ_f , cosmologists can determine if the observed angle is consistent with a flat, open or closed geometry (see Fig. 17.2). It turns out that to a very high accuracy, the universe is flat, in full agreement with the measurements of the energy density of the universe.

As we already mentioned, on large angular scales the temperature fluctuations in the plasma and the gravitational redshift work in opposite directions, so the observed fluctuations are due to the difference between the two effects. On the other hand, in the fundamental peak, the hot and cold plasma regions switch places, while the high and low density dark matter regions do not move. As a result the plasma temperature fluctuations and the gravitational redshift due to dark matter are now added together. This is why the first peak is so high. In the second peak the two effects are again opposite. By comparing the

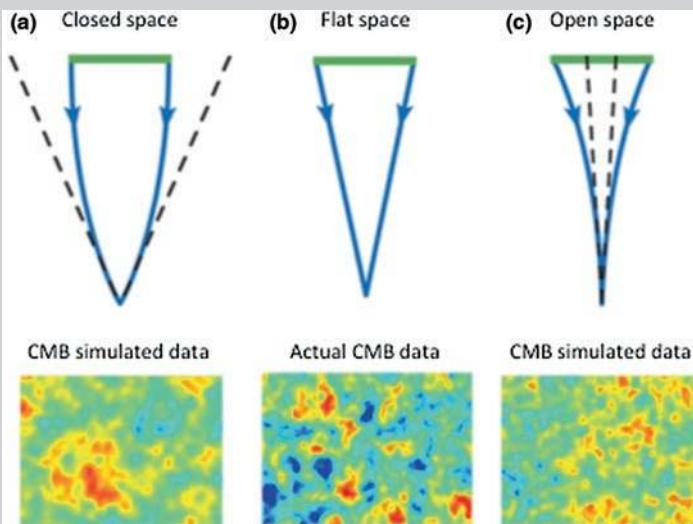


Fig. 17.2 CMB measurements show that the universe is flat. In the top figures, the solid green lines represent the physical size of the fundamental wavelength and the blue lines represent light propagation in closed, flat and open universes. The fundamental wavelength and the light rays emanating from its two ends form a giant triangle in space (as discussed in Chap. 4, the sum of angles in a triangle can be more than or less than 180° depending on the geometry of space). The black dotted lines indicate the angle subtended by the fundamental wavelength for each geometry. In a closed universe the fundamental wavelength subtends a larger angle than in flat space, and thus hot and cold spots would appear larger, as shown in the simulated CMB data in part (a). In an open universe the fundamental wavelength subtends a smaller angle than in flat space, and thus hot and cold spots would appear to be smaller, as shown in the CMB simulation in part (c). The actual CMB data in part (b) is consistent with a flat universe. Credit NASA

heights of the first and second peaks, cosmologists have been able to determine the relative amounts of atomic and dark matter. The results are perfectly consistent with calculations based on nucleosynthesis. Such consistency checks provide reassurance that our understanding of the early universe is on the right track.

The prediction of a scale invariant primordial spectrum is one of inflation's most important features. In addition to the CMB data which favor inflation, numerical simulations have been very successful in reproducing observations

of the large-scale structure of the universe—starting from the primordial scale-invariant spectrum (see Chap. 12).

17.3 Gravitational Waves

Another key prediction is that the tumultuous epoch of inflation generated gravitational waves that should also have a scale invariant spectrum (Fig. 17.3). As we discussed in Chap. 4, Einstein predicted the existence of gravitational waves nearly one hundred years ago. Gravitational waves can be detected because they stretch and squeeze space as they pass through it (without changing the volume), and this can cause distances between objects to change (see Fig. 17.4). However, this is a minuscule effect: for example, if a gravitational wave produced by a close pair of neutron stars rotating about one another were to pass between you and your friend on the other side of the room, the distance between you would be altered by less than the size of a proton! Although extremely challenging, gravitational waves from astrophysical sources have been recently detected (as discussed in Chap. 4).



Fig. 17.3 The Russian physicist Alexei Starobinsky was the first to show that gravitational waves would be generated during an inflationary period. He did so in 1979 in the context of the “Starobinsky model”, which predated Guth’s version of inflation. Credit PR Image iau1304b, Alexei Starobinsky, recipient of the 2013 Gruber Prize (<https://www.iau.org/news/pressreleases/detail/iau1304/>)



Fig. 17.4 A gravitational wave alternately stretches and squeezes a ring of freely floating test particles as it passes by

The origin of gravitational waves from inflation is similar to that of the primordial density fluctuations. The waves are produced by quantum fluctuations in the geometry of space and time. They originate in tiny regions of the Hubble size d_H , and then their wavelengths are stretched to astronomical sizes by the rapid inflationary expansion. The magnitude of the fluctuations is set by the Hubble parameter H , which is in turn determined by the false vacuum energy density [see Eq. (16.2)]. Since H remains nearly constant during inflation, the amplitude of gravitational waves is about the same for all wavelengths. In other words, the predicted gravitational wave spectrum is scale invariant.

Once created, primordial gravitational waves propagate through the universe. The predicted amplitude of the waves is too small to be directly detected with instruments like LIGO. However, primordial gravitational waves are expected to leave an imprint on the CMB radiation, both by impacting temperature fluctuations and by causing specific polarization patterns.

The reason gravitational waves can cause temperature fluctuations is because as the waves pass through the plasma at recombination, in some spots they stretch the plasma in our direction—that is, in the direction where our galaxy will eventually emerge, causing photons from those regions to be somewhat blue-shifted and thus hotter. In other spots the gravitational waves cause regions of plasma to be compressed away from us, and such regions appear red-shifted and thus cooler. It is difficult, however, to distinguish these temperature variations from those caused by primordial density fluctuations. But the polarization patterns induced by gravitational waves have a unique signal.

When photons scatter off electrons in the cosmic plasma, they get polarized—which means that the electric field of the photon gets oriented in a certain way (determined by the direction of motion of the incoming and scattered photons). With a large number of photons undergoing multiple

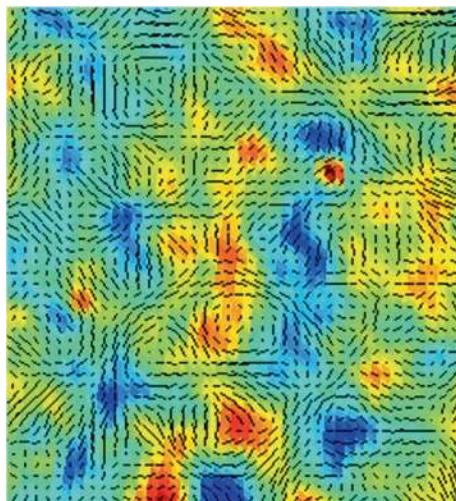


Fig. 17.5 Temperature hot and cold spots, plus polarization (*black line segments*) as measured by the Boomerang detector. The direction of polarization (that is, the direction of the electric field) in a region of the sky is indicated by line segments. *Credit BOOMERanG experiment*

scatterings, there is no net polarization. But at the epoch of recombination, just before the universe became transparent to radiation, the CMB photons scattered for the last time. The photons that we see come mostly from denser plasma regions, and we can see only photons that scattered in our direction. As a result, the observed CMB radiation is polarized (see Fig. 17.5). Primordial density fluctuations produce an E-mode pattern, consisting of radial and ring-like structures (see Fig. 17.6). In addition to E-modes, polarization caused by gravitational waves displays a swirl-like pattern that can be clockwise or anticlockwise; such patterns are called *B-modes*.

Many collaborations around the world have been searching for traces of primordial B-mode polarization. In March 2014, the BICEP 2 team² announced that they had found a pattern of polarization that is consistent with gravitational waves from inflation (see Fig. 17.7). Unfortunately,

²BICEP stands for Background Imaging of Cosmic Extragalactic Polarization.

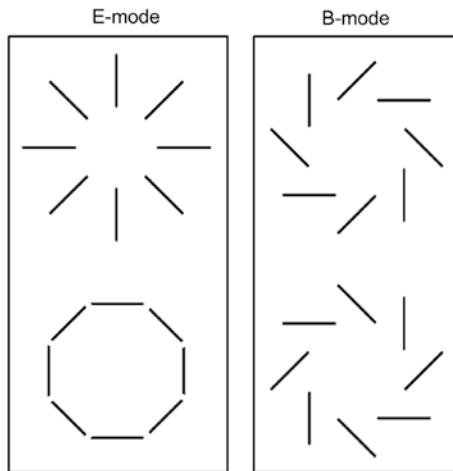


Fig. 17.6 Polarization pattern for E and B-modes. B-modes have a “curl” or a swirl-like pattern, and are produced by primordial gravitational radiation

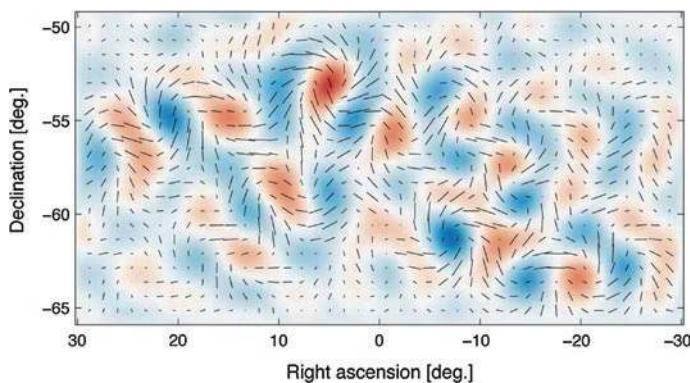


Fig. 17.7 BICEP 2 polarization data. The segments indicate the direction of polarization, after the E-mode pattern has been removed. The blue and red spots indicate whether the B-modes are clockwise or anticlockwise, respectively. Tightly wound spots have higher color intensity. Credit From: BICEP2 collaboration, Detection of B-Mode Polarization at Degree Angular Scales by BICEP2. PRL 112, 241101 (2014)

subsequent analysis and collaboration with other scientific teams showed that their detected B-mode polarization is probably due to galactic dust.

The search for B-modes continues in next generation experiments, and researchers are hopeful that they will detect signs of primordial gravitational

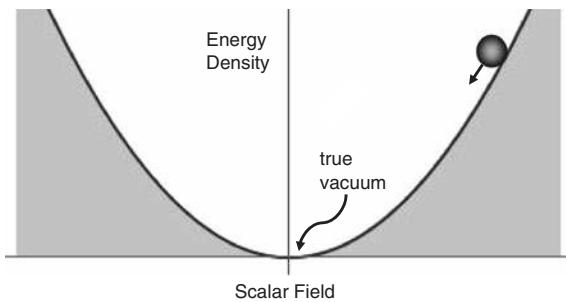


Fig. 17.8 “Topless” inflation model

waves soon. The stakes are very high, for two main reasons: Firstly, inflation predicts that the amplitude of the gravitational waves is proportional to the energy scale at which inflation takes place. Thus, if we can measure the magnitude of gravitational wave perturbations, we stand to learn about the physics behind inflation and about physics at energies that are far too high to be studied in accelerators. And secondly, these gravitational waves are thought to be produced via quantum mechanical effects, thus their existence could shed light on the unification of gravity and quantum mechanics.

17.4 Open Questions

The theory of inflation explains many puzzling features of the big bang and makes observational predictions that have been beautifully confirmed by the data. We thus have good reasons to believe that a period of accelerated inflationary expansion did occur in the early universe. This does not mean, however, that the problem of the origin of the universe has been solved.

First, we should emphasize that inflation is not a specific model, like the Standard Model of particle physics, but rather a paradigm encompassing a wide class of models. The main differences between the models are in the choice of the inflaton potential energy landscape. Linde's 1982 model assumed an energy hill with a flat hilltop, as illustrated in Fig. 16.8. A few years later, Linde proposed another model, where the energy hill keeps rising without limit in both directions (Fig. 17.8). Such a “topless” hill has a true vacuum at the bottom and no definite location for the false vacuum. The role of the false vacuum can be played by some point on the slope, where the inflaton field starts its downward roll. If the slope is sufficiently gentle, the field will roll slowly, and inflation will occur. Yet another possibility is a hybrid of

Linde's hilltop model with Guth's original scenario (we shall discuss this in more detail in Chap. 18). Cosmologists have also studied models of inflation including several scalar fields and models where the inflaton field is incorporated in a particular particle physics theory.³

Despite the variety of models, the predictions of inflation are rather robust. All models predict a nearly flat universe and an almost scale-invariant spectrum of density fluctuations. There are, however, some differences in the details. For example, the predicted small deviations from scale invariance are different for different models. The Planck satellite observations disfavor the "topless" models of Fig. 17.8, and as the empirical data continue to pour in, we can expect a further reduction in the number of viable models.

Even if we converge on a single model of inflation, that will not be the end of the story. All models assume that at the onset of inflation the universe was in a state of false vacuum. Why was it so? Only a tiny nugget of false vacuum is required, but even a small initial nugget calls for an explanation. Where did it come from? We shall discuss this and other questions raised by the theory of inflation in subsequent chapters.

Summary

Inflation makes several predictions, three of which we discussed here:

The universe is flat on the largest observable scales; density fluctuations have a nearly scale-invariant spectrum; and a scale-invariant spectrum of gravitational waves should also be present. The first two predictions have been observationally confirmed and the search for primordial gravity waves is currently underway. The idea of inflation appears to be on the right track and has by now become the leading cosmological paradigm.

Questions

1. Why was the discovery of dark energy such a boost for the theory of inflation?
2. Why are CMB temperature anisotropies so important?

³Alexei Starobinsky suggested a model of inflation without scalar fields. In this model, the accelerated expansion of the universe is due to a quantum modification of Einstein's equations. Starobinsky introduced his model in 1979, before Guth published his first paper on inflation. But he did not realize that an accelerated expansion period explains the puzzling features of the big bang, so Guth is generally credited with the idea of inflation.

3. Why do temperature anisotropies on large angular scales (greater than 2°) represent density fluctuations as they emerged immediately following inflation?
4. Do hot patches in the CMB on large angular scales emerge from under- or over-dense regions? Explain your answer.
5. From the data in Fig. 17.1, estimate the magnitude of CMB temperature anisotropies on large angular scales. Can you tell from the figure that the spectrum of primordial density perturbations is approximately scale invariant?
6. What is a primordial sound wave?
7. Why are sound waves with varying wavelengths present in the early universe?
8. What physical processes give rise to the peaks in Fig. 17.1? Why is the fundamental peak higher than the other peaks?
9. Calculate the angle θ subtended by the horizon distance at recombination. *Hint:* in this calculation you can go through the following steps. First recall from Chap. 7 that the horizon distance at time t in the matter era is $d_{hor}(t) \approx 3ct$. Once you calculated $d_{hor}(t_{rec})$, find its present size, d , by accounting for the expansion of the universe from t_{rec} to present. The angle θ subtended by a distance d on the surface of last scattering can be found from the formula⁴ $\theta = d/d_{ls}$ where the current distance to the surface of last scattering is $d_{ls} \approx 46 \times 10^9$ lyrs. (In Sect. 15.2 of Chap. 15 we explained that, because the CMB photons were last scattered so early in the universe's history, the distance to the surface of last scattering is approximately equal to the present horizon distance).
10. Find the physical size of the fundamental wavelength λ_f . *Hint:* λ_f should be equal to twice the distance traveled by sound from the big bang to the time t_{rec} (because the period of these waves is $2 t_{rec}$). You can use the horizon distance $d_{hor}(t_{rec})$ calculated in Question 9 and the fact that sound waves propagate at 0.6 of the speed of light.
11. Explain how the CMB data is used to measure spatial curvature.
12. How could primordial gravitational waves cause temperature fluctuations in the CMB?
13. What do physicists mean when they say that light is polarized?

⁴This formula assumes a flat geometry and gives the angular size in radians. If you want to express θ in degrees, you can use 2π radians = 360° .

14. If scientists detect polarization in the CMB that is caused by primordial gravitational waves, what information can we learn about the universe?
15. Is inflation a specific theory or is it more like a general framework? In your opinion, is this good or bad? Do you think gravitational wave physics might help to hone in on a specific theory of inflation?
16. Does the theory of inflation fully explain the origin of the universe? If not, what questions does it leave unanswered?

18

Eternal Inflation

Inflation enlarges the size of the universe by an enormous factor, so we can observe only a tiny part of it. The theory explains very well what we see in this small domain, but it also makes predictions about the parts of the universe that we cannot see—beyond our cosmic horizon. This has led to a radical revision of our global view of the universe.

18.1 Volume Growth and Decay

In very general terms, the new worldview can be understood as follows. An inflating universe is governed by two competing processes: exponential growth of false vacuum volume and the decay of false vacuum. This is similar to the reproduction of bacteria which multiply by division and are destroyed by antibodies. The outcome depends on which process is more efficient. If the bacteria are destroyed faster than they reproduce, they will quickly die out. Alternatively, if the reproduction is faster, bacteria will rapidly proliferate. In most models of inflation, the rate of volume expansion is much higher than that of false vacuum decay. This means that expansion wins, and the total volume of inflating regions grows with time.

False vacuum decay is induced by probabilistic quantum processes, so it happens in random locations at random times. The result is a stochastic patchwork of true and false vacuum regions. In Fig. 18.1 we illustrate the dynamics schematically, using a simple 2D model. We start with a false vacuum region, shown as a white square in the first frame of the figure. The following three

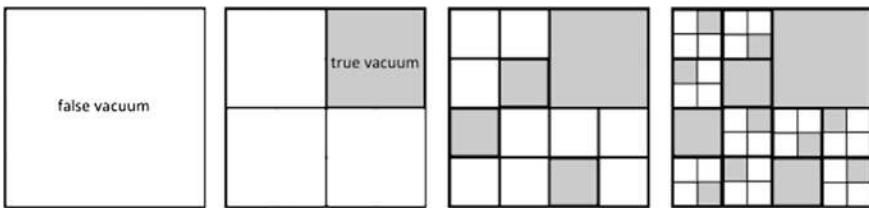


Fig. 18.1 The area of the false vacuum quadruples and one out of four newly created regions immediately decays to the true vacuum state at each time step (from *left* to *right*). True vacuum regions are indicated by grey shading. The smallest squares in each frame have the same physical size, but appear to decrease with time, because the expansion factor is taken out

frames show the same region after three consecutive doubling times. To avoid running out of space in the figure, we use “comoving coordinates”, factoring out the expansion of the universe, so all four “snapshots” of the region have the same apparent size. In the second frame, the size of the region has doubled, and its area has quadrupled, so it now contains four squares of the same physical size as the original one. We assume that false vacuum has decayed into true vacuum in one of the squares, indicated by grey shading. In the third frame, the size of false vacuum regions has doubled again, and one quarter of the false vacuum regions have been converted into true vacuum. The same scenario plays out in the fourth frame.

This simple algorithm can be repeated any number of times. At each time step, the false vacuum area is quadrupled, and one quarter of it is lost to decay. The resulting change in the amount of false vacuum is by a factor of $4 \times \frac{3}{4} = 3$. After N steps this amount will grow by a factor of 3^N . Thus the expansion of false vacuum more than makes up for its loss due to vacuum decay.

If false vacuum regions multiply faster than they decay, inflation never ends in the entire universe. Even though it ended in our local region, it still continues in remote parts of the universe, producing new true vacuum regions like ours. This never ending process is called *eternal inflation*.

Fractals

The pattern of true and false vacuum regions obtained by repeated application of the algorithm in Fig. 18.1 is an example of what mathematicians call a self-similar fractal. “Self-similar” refers to the fact that the pattern is statistically the same on every distance scale. If, for example, we pick a small white square in the last frame, representing a false vacuum region, its subsequent evolution will be essentially the same as that of the initial white square in the first frame.

The evolution will not be exactly the same, because there is an element of randomness in the algorithm. But after many steps the statistical properties of the regions will be very similar.

The term “fractal” refers to the fact that the inflating part of space in this model has, in a certain sense, a fractional dimension. If you double the size of a one-dimensional line, its length will increase by a factor of 2. If you double the size of a 2D figure, its area will increase by a factor of $2^2 = 4$. And if you double the size of a 3D body, its volume will increase by a factor of $2^3 = 8$. In general, when the size of a d -dimensional object is doubled, the amount of “stuff” in the object is increased by a factor of 2^d . Now, in the model of Fig. 18.1, the area of the inflating part of space grows by a factor of 3 in one doubling time. This is between $2^1 = 2$ and $2^2 = 4$, suggesting that the fractal dimension of the inflating region is between 1 and 2. To find the exact dimension, we have to solve the equation $2^d = 3$. The solution is $d = \log_2 3 = 1.58$.

The eternal nature of inflation was first recognized by Vilenkin in 1983, soon after Guth proposed his theory of cosmic inflation, and was later investigated by a number of physicists, most notably by Linde. Inflation is eternal in nearly all models that have been studied so far. It is possible to construct non-eternal models, but they require rather contrived potential energy landscapes for the inflaton scalar field.

The simple model of Fig. 18.1 captures only rough features of an eternally inflating universe on very large distance scales; the details depend on the specific mechanism of false vacuum decay. There are two such mechanisms to consider: quantum random walk and bubble nucleation. We shall now discuss them in turn.

18.2 Random Walk of the Inflaton Field

As we discussed in Chap. 16, small density fluctuations are generated during inflation, because the inflaton scalar field is subjected to random quantum kicks as it rolls down the potential energy hill (see Fig. 18.2). While the field is rolling downwards, the quantum kicks are much weaker than the force due to the slope of the hill, and that is why the field reaches the bottom everywhere at about the same time, yielding only small density fluctuations.

But now let us ask ourselves: What happens when the field is close to the top of the hill, where the slope is very small? There, the inflaton is at the mercy of quantum kicks, which shove it randomly one way and then the other. The typical time interval between the kicks is the doubling time of inflation, $t_D \sim 1/H$ (recall H is the Hubble parameter); hence the field will undergo a “random walk”, making random steps forward and backwards,

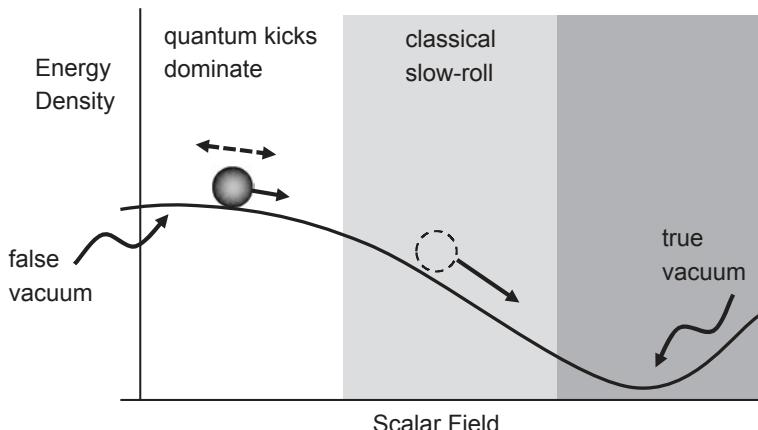


Fig. 18.2 At the top of the hill quantum kicks (long dashed arrows) are stronger than classical motion (small solid arrow), so the field undergoes a random walk. Once the slope is steep enough, the classical motion dominates (see long solid arrow), and the field slow-rolls to the end of inflation. This classical slow-roll regime is highlighted in light grey

separated by time intervals $\sim 1/H$. Eventually, after a number of steps, the field will get to a steeper part of the hill and then roll down towards the end of inflation.¹

To see how the inflaton field values are distributed in space, let us recall that quantum kicks occur in small patches of Hubble size $d_H = c/H$. This is the maximum distance over which communication is possible in the inflating universe, so the directions of the kicks in different “Hubble patches” are random and independent of one another. If two points in space are separated by less than a Hubble distance, they experience the same quantum kicks. But the points are rapidly driven apart by the inflationary expansion, and once their separation exceeds d_H , their histories begin to diverge. As time goes on, the distance between the points gets larger and larger, and the field values become more and more divergent.

The smallness of density fluctuations in our observable region tells us that all points within our region were still within a Hubble distance of one

¹In a “topless” model, having a potential energy landscape like the one shown in Fig. 17.8, the hill gets steeper at higher altitudes, so the classical force pushing the inflaton field downwards gets stronger. But Andrei Linde has shown that the strength of quantum kicks increases with altitude even faster. Thus, if the inflaton field starts out at high enough elevation, quantum kicks become the dominant force, and the field undergoes a quantum random walk, until it gets to a sufficiently low level and rolls classically downhill.



Fig. 18.3 2D simulation of an eternally inflating spacetime (performed by V. Vanchurin, A. Vilenkin and S. Winitzki). It shows true vacuum islands (*light*) in the inflating background (*dark*). The larger islands are the older ones: they have had more time to grow (*Note* that the color coding is different here than in Fig. 18.1)

another when the inflaton field was well on its way down the hill. That is why the effect of quantum kicks was very minor, and the field reached the bottom everywhere at about the same time. But if we could go to very large distances, far beyond our horizon, we would see regions that parted our company when the field was still wandering near the hilltop. Such regions have very different scalar field histories, and some of them may still be in the process of inflationary expansion.

Eternally inflating spacetimes produced via a quantum random walk have been studied in various computer simulations. Figure 18.3 is a snapshot of a 2D simulation which shows that true vacuum regions form as islands in the inflating background of false vacuum. The islands grow rapidly in size, as their boundaries advance into the inflating sea, but the inflating regions that separate them expand even faster, making room for more islands to form. The resulting pattern resembles an aerial view of an archipelago, with large islands surrounded by smaller ones, which are surrounded by still smaller ones, and so on. This fractal pattern is somewhat similar to that in our simple model of Fig. 18.1; the main difference is that the islands do not have orderly square shapes and are distributed in a more irregular manner.

18.3 Eternal Inflation via Bubble Nucleation

Suppose now that the false vacuum is separated from the true vacuum by an energy barrier, as in Guth's model of inflation (see Fig. 16.6). Then the false vacuum decays through bubble nucleation. Bubbles of true vacuum pop out at random here and there and immediately start to expand. They expand faster and faster, approaching the speed of light, but they are driven apart by the expansion of intervening regions of false vacuum. Hence, in this scenario inflation never ends—it is eternal. This was bad news for Guth's original model because it was unclear how the false vacuum energy could ever be turned into a hot fireball. But later Paul Steinhardt realized that this could be achieved by modifying the shape of the inflaton energy landscape. Instead of a steep decline towards the true vacuum, he suggested that the barrier should be followed by a gentle slope, as in Fig. 18.4. Then the inflaton field in a newly formed bubble has a value on the right hand side of the barrier; this value is separated from the true vacuum by a long stretch of gentle slope.

While the bubble expands, inflation continues inside of it, as the field slowly rolls downhill. When the field gets to the bottom of the hill, it converts its energy into a hot fireball of particles. This model is thus a hybrid of Guth's original scenario and Linde's slow roll model. The false vacuum inflates eternally, producing an unlimited number of bubbles, and each of the bubbles undergoes a period of slow-roll inflation in its interior, followed by the production of a fireball and subsequent hot big bang evolution. If we could take a "bird's-eye view" of the eternally inflating false vacuum with bubbles, the picture would be the same as in Fig. 16.7, except now inflation continues within each bubble, as the field slowly rolls towards the true vacuum.

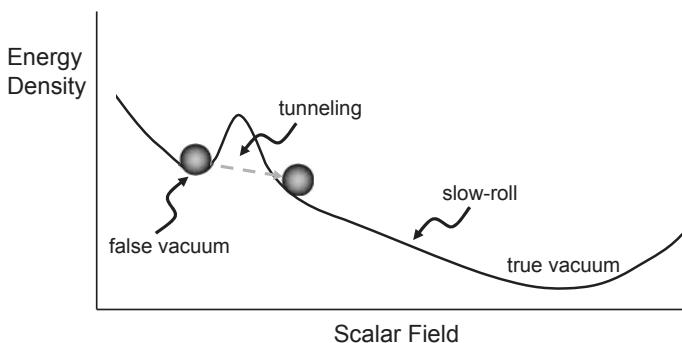


Fig. 18.4 Eternal inflation via bubble nucleation

According to this scenario, we live in one of the bubbles and can see only a small part of it. No matter how fast we travel, we cannot catch up with the expanding boundaries of our bubble. So, apart from rare bubble collisions, for all practical purposes, each bubble is a self-contained, isolated bubble universe.

18.4 Bubble Spacetimes

Bubble universes have a very interesting spacetime structure, which we shall now discuss in detail. Bubbles are microscopic when they materialize; then they expand without bound and become arbitrarily large. The central parts of large bubbles are very old. They evolved through all the phases of the hot big bang. Stars formed and died, intelligent life emerged and went extinct, so now these old regions are dark and barren. On the other hand, regions at the bubble periphery are young. This is where the false vacuum energy is being converted into a hot fireball and new stars are being formed.

The spacetime of a bubble universe is schematically illustrated in Fig. 18.5. Here, the vertical direction is time, the horizontal direction is space, and two of the three spatial dimensions are not shown. Each horizon-

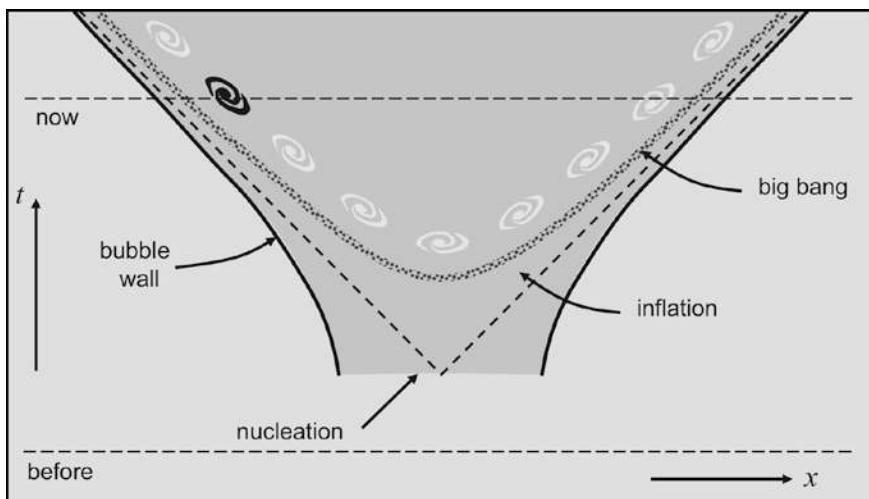


Fig. 18.5 Spacetime diagram of bubble universe (showing one spatial dimension only). The dashed lines at 45° angles are light signals sent outwards from the bubble center at the time of bubble formation. Note that these signals never catch up with the expanding boundaries of the bubble

tal slice through the diagram gives a snapshot of the universe at a moment of time. You can follow the history of the bubble by starting with the horizontal line marked “before” at the bottom of the figure and gradually moving it upward. The horizontal segment marked “nucleation” indicates the moment of bubble formation. The fuzzy grey line shows where the fireball is formed and hot big bang evolution begins. The location marked by a black galaxy is the here and now, and white galaxies indicate spacetime regions where the conditions are similar to what we have here today. The horizontal dashed line labeled “now” represents the present time. It shows the bubble universe with a barren central region and some hot evolving regions close to the boundaries.

There is, however, another way to think about this spacetime, which yields a very different view of the bubble universe. The key point is that “a moment of time” is not a uniquely defined concept in general relativity. When cosmologists talk about a moment of time, they picture a large number of observers, equipped with clocks and scattered through the universe. Each observer can see only a small region in her immediate vicinity, but the whole assembly of observers is needed to describe the entire spacetime. We can think of ourselves as one member in this assembly. Our clock now shows the time 13.8 billion years ABB. “The same time” in another part of the universe is when the clock of the observer located there shows the same reading. We have to decide, though, how observers, who are outside each other’s horizons, are to synchronize their clocks.

In the case of a Friedmann universe, the answer is simple: the big bang is the natural origin of time, so each observer should count time starting from the big bang. But in an eternally inflating spacetime with multiple bubble universes, there is no such obvious choice. One possibility is to imagine observers who can exist in false vacuum and who synchronize their clocks in a small false vacuum region, while they are still within each other’s Hubble distance. The observers are then driven apart by the inflationary expansion and encompass a large volume, including many bubble universes, at later times. The snapshots of the eternally inflating universe in Figs. 16.7 and 18.3 assume such a group of observers, and the moments “before” and “now” in Fig. 18.5 correspond to this choice as well.

But suppose now that we want to describe a specific bubble universe from the point of view of its inhabitants. Then the situation is similar to that of a Friedmann universe: there is now a natural choice for the origin of time. All observers inhabiting the bubble universe can count time from their local “big bang”, that is, from the creation of the fireball at their respective locations. To distinguish between the large-region and single-bubble description

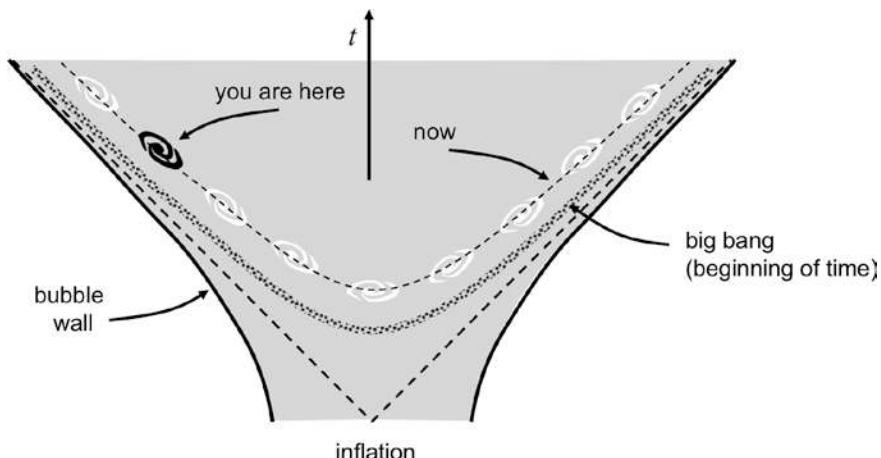


Fig. 18.6 Viewed from the inside (internal view), each bubble is an infinite universe. In the global view of Fig. 18.5, each bubble can grow for an infinite time, but is finite at any given moment of time. The difference is due to different definitions of time

we shall refer to them as “global” and “local” (or “internal”) views, respectively.

The internal view of the bubble universe is illustrated in the spacetime diagram of Fig. 18.6. The spacetime structure is the same as in Fig. 18.5, but the lines representing moments of time are drawn differently. The fuzzy grey line representing the creation of the fireball now corresponds to the initial moment. The density of matter at this moment is very nearly uniform, and thus in the local view the bubble universe is nearly homogeneous (apart from small inhomogeneities due to quantum fluctuations). The present moment in this view is represented by the dotted line marked “now”, which coincides with the line of galaxies in the figure. All points on this line are characterized by the same density of matter and the same average density of stars as observed in our local region. But most remarkably, from the local point of view the bubble universe is infinite.

In the global view, the bubble universe grows with time, as new hot fireball regions are created near its boundary, and becomes arbitrarily large if you wait long enough. But in the local view, the fireball is created all at once and the bubble universe is infinite from the very beginning. In Fig. 18.6 this infinity is evident from the fact that the fuzzy line representing the creation of the fireball never comes to an end. Analysis shows that the spatial geometry of a bubble universe in the local view is that of an open (negative

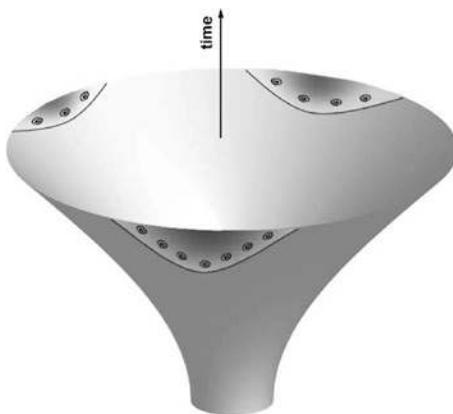


Fig. 18.7 The 2D surface represents the spacetime of a one-dimensional universe. This universe is closed and finite. It is filled with false vacuum at the initial moment (*bottom* of the figure) and contains three bubble universes at the time corresponding to the top of the figure. Each bubble universe appears to be infinite from the point of view of its inhabitants

curvature) Friedmann universe.² Thus, the picture of finite spatial sections which grow for an infinity of time in the global viewpoint is replaced by an infinity of spatial extent at each moment of time in the internal viewpoint.

This dual viewpoint leads to a very interesting situation: an eternally inflating spacetime can be closed and finite, and yet it can contain bubble universes that appear to be infinite to the observers who live inside (Fig. 18.7).

We note finally that analysis of random walk models of eternal inflation has shown that the properties of true vacuum islands predicted in these models (and illustrated in Fig. 18.3) are similar to those of bubble universes. An island also appears to be infinite to internal observers, and the observers cannot escape from their island, because its boundaries are expanding so fast.

²By the end of inflation the bubble universe becomes nearly flat, so its curvature is very difficult to observe.

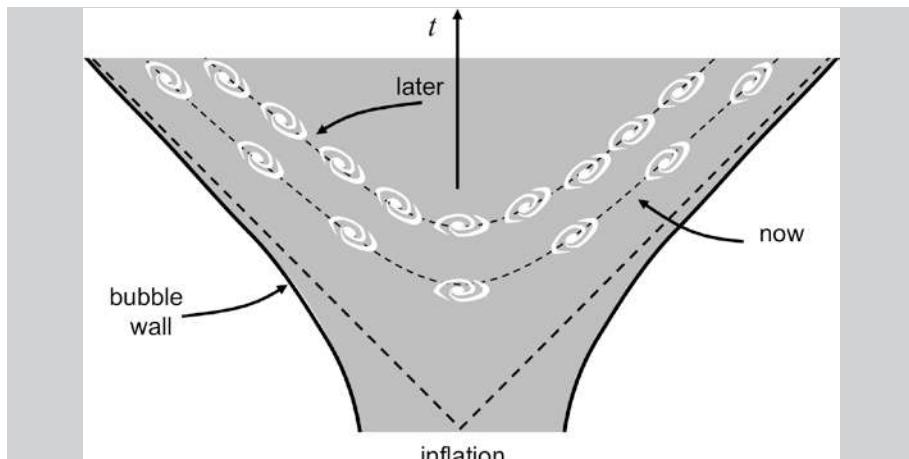


Fig. 18.8 A contracting bubble universe. The galaxies get closer together, even though the bubble radius grows with time

More on bubble spacetimes

Observers can find the expansion rate of their bubble universe by measuring how fast the distances between galaxies grow with time. But because of the complicated spacetime geometry, this rate is not simply related to, and is typically much slower than, the expansion rate of the false vacuum outside. It is even possible for the bubble interior to contract, while the bubble itself is expanding. An external observer would then see the bubble radius grow, while internal observers would see galaxies getting closer with time, as illustrated in Fig. 18.8. This situation could arise if the vacuum energy density (the cosmological constant) inside the bubble is negative. A negative cosmological constant produces an attractive gravitational force and causes the bubble universe to contract to a big crunch.

18.5 Cosmic Clones

At this point we would like to mention a remarkable and, to our minds, disturbing consequence of eternal inflation. Because the number of bubble universes is unlimited, and each of them expands without bound, they will contain an unbounded number of regions with the same size as our observable universe; let us call them *O*-regions. All these regions look the same at the end of inflation, except for the pattern of small density fluctuations. As fluctuations are amplified by gravity, the properties of the regions diverge,

and they end up with different distributions of stars and galaxies. Random quantum events also influence the evolution of life, and this leads to further divergence of histories. So, should we expect the infinite number of *O*-regions to each have unique histories that result in their own unique present state?

In classical physics, the state of a physical system is described by specifying the precise positions and velocities of all its particles. Given a system of particles—say the contents of your refrigerator right now—you can always change its state by an arbitrarily small amount. Even if you barely changed the position or velocity of one single particle in the milk bottle, you would create a new distinct state. Classically, there is a continuum of states, which can be made arbitrarily close to one another, yet still maintain a unique identity. In quantum mechanics this is impossible because the uncertainty principle leads to an inherent fuzziness in the state of a system. Configurations which are too close to one another cannot be distinguished, even in principle. The upshot of the uncertainty principle is that the number of distinct quantum states in any finite volume is finite.

The number of possible histories of an *O*-region is finite as well. A history is described by a sequence of states at successive moments of time. The histories that are possible in quantum physics differ immensely from the ones possible in the classical world. In the quantum world the future is not uniquely determined by the past; the same initial state can lead to a multitude of different outcomes, and so only the probabilities of those outcomes can be determined. Consequently, the range of possible histories is greatly enlarged. Once again, though, the fuzziness imposed by quantum uncertainty makes it impossible to distinguish histories that are too close to each other. An estimate of the number of distinct histories that can unfold in an *O*-region between the big bang and the present gives $\sim 10^{10^{150}}$. This number is fantastically huge, but the important point is that the number is finite.

Let us now take stock of the situation. The theory of inflation tells us that the number of *O*-regions in an eternally inflating universe is infinite, and quantum uncertainty implies that only a finite number of histories can unfold in any *O*-region. The initial states of the *O*-regions at the big bang are set by random quantum processes during inflation, so all possible initial states are represented in the ensemble. Putting those statements together, it follows that every history which has a nonzero probability should be repeated an infinite number of times.

Among the infinitely replayed scripts are some very bizarre histories. For example, a huge quantum fluctuation could cause the Sun to suddenly collapse to a black hole. The probability of this happening is extremely small,

but remember: in quantum mechanics all processes that are not strictly forbidden by conservation laws do occur with a nonzero probability.

A striking consequence of this picture of the world is that there should be an infinity of O -regions with histories absolutely identical to ours. That's right, scores of your duplicates are scattered throughout the eternally inflating spacetime. They live on planets exactly like Earth, with all its mountains, cities, trees, and butterflies. There should also be regions where histories are somewhat different from ours, with all possible variations. For example, some readers will be pleased to know that there are infinitely many O -regions where Hillary Clinton is the President of the United States.

You may be wondering whether all these things in different regions are happening at the same time. This question does not have a definite answer, because time and simultaneity are not uniquely defined in general relativity (as we discussed in Sect. 18.4). If, for example, we use the local time definition in a bubble universe, then at each moment of time the bubble interior is an infinite hyperbolic space, and each of us has an infinite number of duplicates presently living in our bubble.³

Note that infinity of space (or time) is not by itself sufficient to warrant these conclusions. We could, for example, have the same galaxy endlessly repeated in an infinite space. So we need some "randomizer", a stochastic mechanism that picks initial states for different regions from the set of all possible states. Even then, the entire set may not be exhausted if the total number of states is infinite. So the finiteness of the number of states N is important for the argument. In the case of eternal inflation, the finiteness of N and the randomness of initial conditions are both guaranteed by quantum mechanics.

18.6 The Multiverse

So far we have assumed that all other bubble universes are similar to ours in terms of their physical properties, but this does not have to be so. Consider for example the energy landscape shown in Fig. 18.9. It has four vacuum states, labeled A, B, C and D, with A having the highest energy density. Vacuum D has the lowest energy density, which is negative in this example.

³If you want to meet some of your duplicates, there is a problem: your nearest cosmic clone lives about 10^{1090} m away. Another issue is that clones who are identical at this time will not remain so, because their subsequent evolution is influenced by random quantum processes.

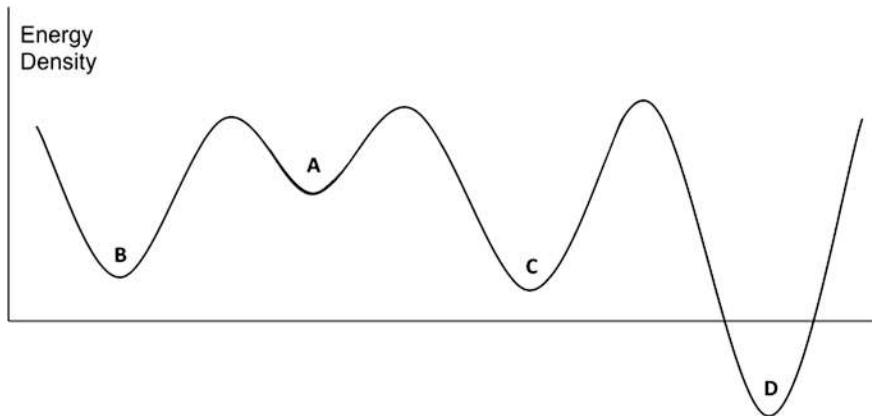


Fig. 18.9 A model energy landscape with four vacuum states, labeled A, B, C and D. Bubbles of B and C can form in vacuum A by quantum tunneling through energy barriers. Similarly, bubbles of D can nucleate in C

Suppose the universe is initially filled with vacuum A. The high energy density of A will then drive exponential inflationary expansion, and bubbles of vacuum B and vacuum C will nucleate and expand in the background of A.⁴ Both B and C have positive energy densities, so the interiors of these bubbles will also be inflating (but at a slower rate than A). Bubbles of vacuum D will nucleate inside the inflating bubbles of C. Furthermore, tunneling “up” from low to high energy density is also possible, albeit with a very low probability—much less than the probability to tunnel “down”. Hence, new bubbles of A will form inside the bubbles of B and C, but they will be very rare.⁵ All of these tunneling processes populate the inflating universe with all four types of vacua (see Fig. 18.10). The number of bubbles of all types will grow without bound in the course of eternal inflation.

A more realistic energy landscape would include several scalar fields. The Higgs field of the Standard Model is an example of a scalar field that we know exists. Grand unified theories predict a number of other Higgs fields whose values determine the particle properties. A model with two scalar fields would have a two-dimensional energy landscape with mountains

⁴We assume here that tunneling from a given vacuum is possible only to a neighboring vacuum in the landscape; hence it is not possible to tunnel from A to D.

⁵Tunneling up from zero and negative energy vacuum states is impossible. We note, however, that such tunneling may occur from zero or negative energy bubbles, if they have inflation or matter dominated periods at the early stages of their evolution (which temporarily increases their overall energy density above the vacuum value of zero or less).

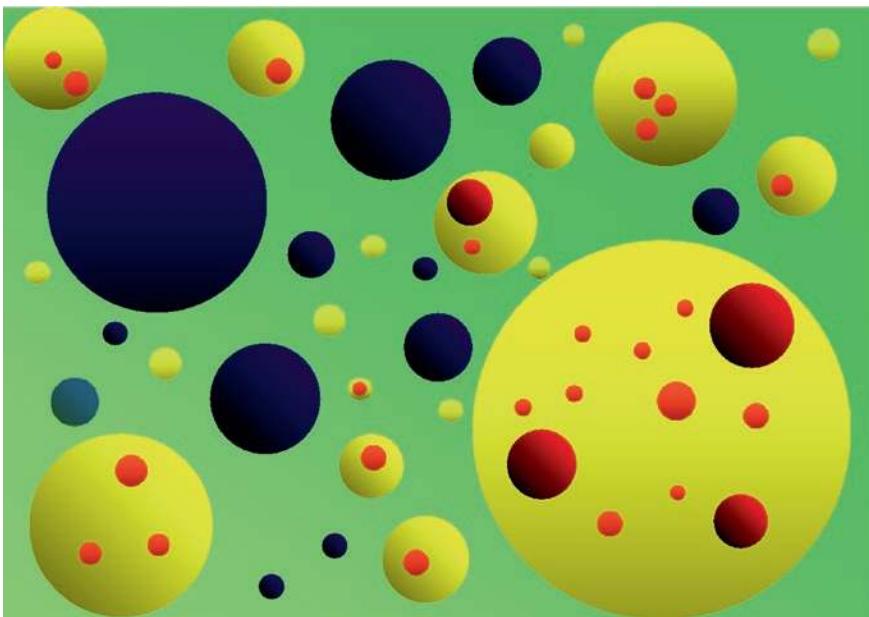


Fig. 18.10 The multiverse of the model with energy landscape shown in Fig. 18.9. Green, blue, yellow and red bubbles correspond to vacua A, B, C and D, respectively. Positive-energy vacua (A, B and C) are inflating, while bubbles of vacuum D do not inflate. Interiors of such negative-energy bubbles eventually collapse to a big crunch

and valleys, as depicted in Fig. 18.11. As before, each valley corresponds to a classically stable vacuum, and transitions between the vacua can occur through bubble nucleation.

With n scalar fields, the energy landscape is n -dimensional. For $n > 2$, we cannot draw such a landscape on a piece of paper, but it is not difficult to analyze mathematically and find all classically stable vacua. As long as inflation begins in one of the positive-energy vacua, all of the other vacua will be realized via the dynamics of eternal inflation. Bubbles of positive-energy vacua will inflate, allowing for more bubbles within bubbles to form, exactly like in the eternal inflation scenario with a single scalar field. Negative-energy bubbles will expand externally, but internally they will contract to a big crunch (see the box in Sect. 18.4).

The values of the Higgs fields vary from one vacuum to another, and as a result particle masses and interactions vary as well. One vacuum state in the energy landscape should correspond to our world, but others are likely to be very different. We thus arrive at the picture of an inflationary *multiverse*, populated by bubble universes with diverse properties.



Fig. 18.11 Energy landscape in a model with two scalar fields. The height represents the value of the potential energy density, and the two axes represent two different scalar fields. Each valley represents a vacuum state. We shall see in Chap. 19 that some modern particle theories predict a large number of such valleys

18.7 Testing the Multiverse

The theory of eternal inflation is mainly concerned with far-away regions, outside our cosmic horizon, and some physicists have raised doubts that the theory can ever be tested observationally. Surprisingly, such tests may in fact be possible.

18.7.1 Bubble Collisions

If a new bubble nucleates within a distance d_H from our expanding bubble, then it will crash into ours. The collision would produce a round spot of higher radiation intensity in the cosmic background radiation. Detection of such spots with the predicted intensity profile would provide direct evidence for the existence of other bubble universes (Fig. 18.12).

The expected number of collision spots in the CMB depends on the rate of bubble nucleation in the false vacuum, and their brightness depends on

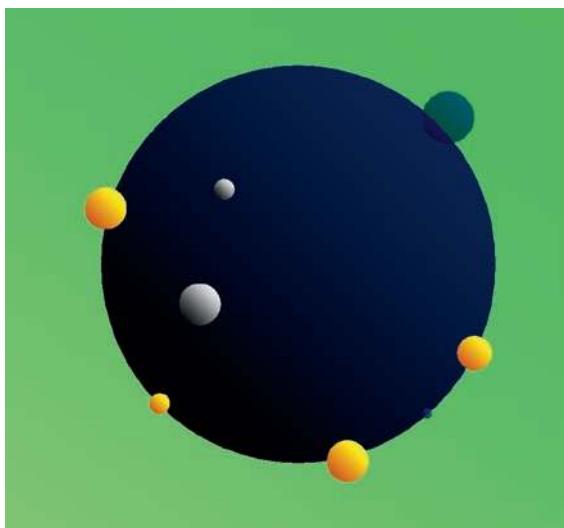


Fig. 18.12 Even though bubble collisions are rare, our expanding bubble will collide with an infinite number of other bubbles in the course of its history

the amount of inflation that took place in our bubble interior: the more inflation, the dimmer are the spots. Unfortunately there is no guarantee that detectable bubble collisions have occurred within our horizon.

18.7.2 Black Holes from the Multiverse

Another interesting possibility is that evidence for the multiverse may be found in our own neighborhood, in the form of black holes. During the slow roll inflation period within our bubble, bubbles of other kinds can nucleate and expand within it. When inflation ends, these bubbles suddenly find themselves surrounded by the very low-energy vacuum that we live in now. At this point they stop expanding and start contracting. (Bubbles expand when they are surrounded by a higher-energy vacuum and contract when the vacuum outside has lower energy.) There is nothing to stop this contraction, so the bubbles collapse to form black holes.⁶

⁶Even though bubbles collapse as viewed from outside, their interiors are filled with a high-energy vacuum and continue to inflate. In a two-dimensional analogy, the resulting geometry can be pictured as an inflating balloon, which is connected to a flat exterior region by a thin “throat”. The throat is seen as a black hole from outside. Thus, black holes formed in this way contain inflating universes inside.

Bubbles that formed earlier have bigger sizes, and bubbles that formed near the end of inflation are very tiny. Bigger bubbles form larger black holes, so the result is a population of black holes with a wide distribution of masses, ranging from less than a gram to millions of Solar masses. These black holes are fossils of the multiverse. If black holes with the predicted mass distribution are discovered, this would provide evidence for the existence of a multiverse.

Apart from these direct methods, some indirect tests of the multiverse theory may also be possible. In fact, some indirect evidence for the multiverse has already been found, as we shall discuss in the following chapters.

Summary

The end of inflation is triggered by quantum, probabilistic processes and does not occur everywhere at once. In our cosmic neighborhood, inflation ended 13.8 billion years ago, but it still continues in remote parts of the universe, where other “normal” regions like ours are constantly being formed.

In the “bubble nucleation” picture, the new regions appear as tiny, microscopic bubbles and immediately start to grow. The bubbles keep growing without bound; all the while they are driven apart by the inflationary expansion of the parent false vacuum, making room for more bubbles to form. We live in one of these bubbles and can observe only a small part of it. The “quantum random walk” picture is similar, giving rise to an infinite number of self-contained island universes separated by inflating false vacuum. Both of these pictures result in a never-ending process called *eternal inflation*. All that is needed for cosmic inflation to be eternal is a false vacuum region that multiplies faster than it decays.

Modern particle physics suggests that a number of Higgs scalar fields should contribute to an energy landscape replete with mountains and valleys. Each valley corresponds to a classically stable vacuum, but transitions between the vacua can occur through bubble nucleation. Thus, if the universe starts in a positive energy vacuum state, then through a random series of transitions from one vacuum to another, all the other vacua in the landscape can be realized. One vacuum state should correspond to our world, but others are likely to be very different. We thus arrive at the picture of an inflationary *multiverse*, populated by bubble universes with diverse properties.

A collision of our expanding bubble with another bubble would produce a round spot of higher radiation intensity in the cosmic background

radiation. A detection of such a spot with the predicted intensity profile would provide direct evidence for the existence of other bubble universes.

An unsettling consequence of eternal inflation is that anything that can possibly happen *will* happen, and it will happen an infinite number of times. In particular, there should be an infinite number of regions absolutely identical to ours. There should also be regions somewhat different from ours, with all possible variations.

Questions

1. What do we mean by the phrase “eternal inflation”? Does it mean that inflation never ends at any given place? Does it mean that inflation continues forever to the past, as well as to the future?
2. Imagine you are a comoving observer in the inflating region represented by one of the white squares in the simple model of Sect. 18.1. What is the probability that inflation will continue in your neighborhood for another doubling time? What is the probability for it to continue for 10 doubling times?
3. Suppose that in each doubling time of inflation the volume of false vacuum grows by a factor $2^3 = 8$ and a fraction f of this volume decays to true vacuum. By how much will the false vacuum volume change after N doubling times? How large should f be in order to prevent eternal inflation from happening?
4. Once a bubble nucleates in an inflating false vacuum, is it possible for inflation to continue inside the bubble?
5. During the course of eternal inflation, how many bubble universes will be formed?
6. Can we in principle travel across the inflating false vacuum and visit other bubble universes?
7. From the external viewpoint, are bubble universes infinite or finite in spatial extent? What about from an internal viewpoint?
8. How is it possible to have a closed and finite universe, which nevertheless contains bubble universes that are infinite from the viewpoint of their inhabitants?
9. Eternal inflation can also be achieved via a quantum random walk. Explain how this works.
10. How would you modify the shape of the potential energy hill in Fig. 18.2 to prevent eternal inflation from happening, while still keeping inflation in the slow-roll region?

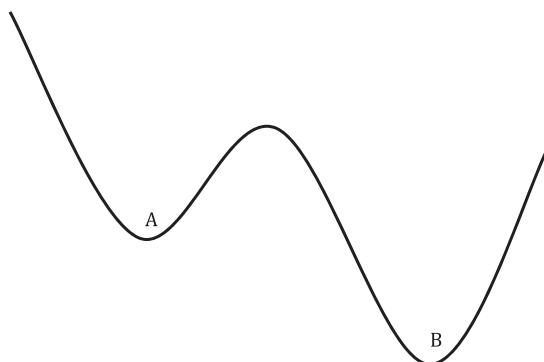


Fig. 18.13 Potential energy density curve with two minima

11. What do we mean by a “multiverse”?
12. Consider the potential shown in Fig. 18.13. Will this model give rise to a multiverse? Assuming that both vacua have positive energy density, sketch a 2-dimensional cartoon of the resulting pattern of bubbles (analogous to Fig. 18.10).
13. Is it possible to test the existence of other bubble universes observationally? If yes, how?
14. Consider the following statement: “In an infinite universe, anything that can possibly happen will happen an infinite number of times.” Is this necessarily true? If not, what additional assumptions about the properties of the universe should we make in order for it to be true?
15. Is the number of distinct states in an infinite volume finite or infinite? Explain. *Hint:* consider an infinite sequence of regions, each of which can only be in one of two states, labeled by 1 and 2. Now consider how many different sequences (consisting of the numbers 1 and 2) are possible.
16. If an eternally inflating universe produces an infinite number of regions having the size of our observable region, and if each region can only have a finite number of histories, is it likely, unavoidable, or impossible to have other regions where someone has had the exact same past as you? If such a person does exist, will she or he have the same future as you?
17. Suppose a region can be in an infinite number of states, and the universe contains an infinite number of such regions. Can we conclude that all possible states will occur somewhere in the universe? (*Hint:* suppose we label the possible states by integers 1, 2, 3, Can you think of an infinite sequence of integers which does not include all possible integers?)

18. Is it possible for the Earth to heat up by suddenly ejecting a huge chunk of ice? Is it likely?
19. Suppose astronomers do not find any signatures of bubble collisions in the CMB. Would that mean that the theory of eternal inflation is wrong? If not, should we still believe that inflation is eternal?
20. How do you feel about the existence of identical Earths? Are you disappointed that our civilization may not be unique? If there is an infinity of other Earths, our civilization appears to be totally insignificant on the cosmic scale. Do you find this upsetting?

19

String Theory and the Multiverse

Much of the research in particle physics has been inspired by the quest for a unified, fundamental theory of Nature. The hope is that beneath the plurality of particles and forces, there is a single mathematical law that governs all natural phenomena. A major step towards the unification of forces was the development of the electroweak theory. The electromagnetic and weak nuclear forces are indistinguishable at very high energies, but at energies below 100 GeV the symmetry between the forces is broken and the two interactions become distinct. In the 1970s and 80s physicists used a similar approach to include the strong nuclear force. They postulated a large, “grand unified” symmetry, which encompasses electroweak and strong interactions and gets broken at very high energies $\sim 10^{16}$ GeV. Grand unification is a very attractive idea, and many physicists believe that it will survive as part of the final theory. However, it suffers from significant shortcomings. First, there is a large (in fact, infinite) number of possible grand unified symmetries to choose from, and none of these symmetries appears to be *a priori* preferred. The list of particles included in the theory is also largely arbitrary. Hence, there is a large number of candidate grand unified theories. This is a problem, since one expects the fundamental theory of Nature to be in some sense unique. Moreover, all attempts to include gravity into the grand unification scheme have proved to be unsuccessful. This led physicists to consider a radically new approach—string theory—which we shall now discuss.

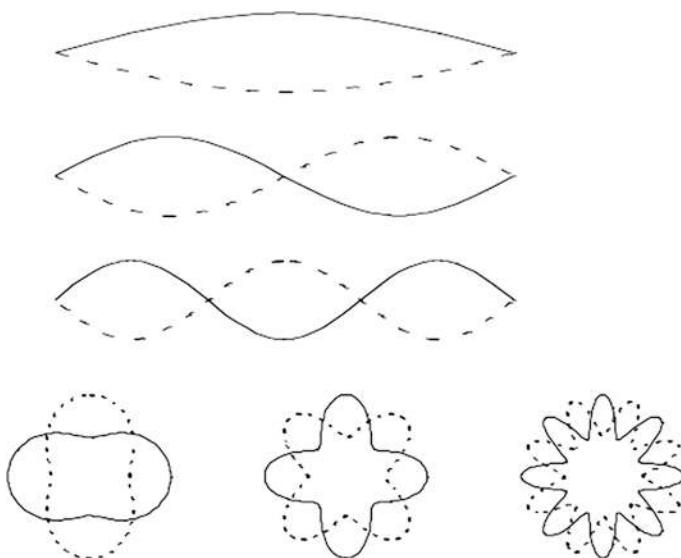


Fig. 19.1 Top three rows a violin string, with its two ends fixed, has a multitude of vibration modes. Bottom row closed strings can oscillate with different modes giving rise to different particles. Strings can also be open, having two free ends. Here, we only consider closed strings for simplicity

19.1 What Is String Theory?

String theory asserts that the basic building blocks of matter are one-dimensional strings, instead of point-like particles. The strings have high tension, which causes them to vibrate at speeds close to the speed of light. All particles of the Standard Model, like electrons or quarks, and any particles not yet discovered, are postulated to be tiny vibrating strings. They appear to be point-like because the strings are so small.

The properties of a particle—its mass, spin, electric and color charges—are determined by the vibration pattern of the string. While each type of particle is made from the same entity—the string—the many possible vibration patterns give rise to a variety of distinct particles. This is analogous to how a single violin string can generate many different musical notes (see Fig. 19.1). Remarkably, one of the possible string vibration patterns has properties that match the graviton—the quantum of the gravitational field. The graviton plays a role in gravity similar to that of the photon in electromagnetic theory. Thus, the problem of unifying gravity with other interac-

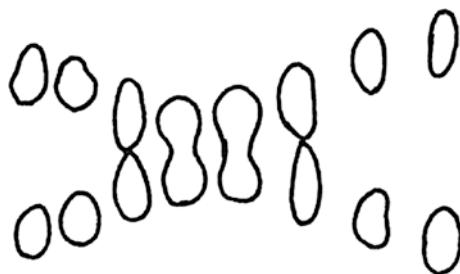


Fig. 19.2 Two strings collide and merge into a single string, which then splits into two again. This corresponds to particle interactions like absorption and re-emission of a photon by an electron

tions does not exist in string theory; in fact, the theory cannot be formulated without gravity.¹

The typical length of vibrating strings is set by the so-called Planck length,

$$\ell_p = \sqrt{\frac{\hbar G}{c^3}} \sim 10^{-35} \text{ m} \quad (19.1)$$

which was introduced by Max Planck at the turn of the 20th century, long before the invention of string theory. Planck realized that ℓ_p is the only quantity with dimension of length that can be constructed out of the fundamental constants G , c and \hbar . It is also the length scale at which quantum fluctuations of spacetime geometry become important, as we shall discuss in the next chapter. The Planck length is incredibly small: it is 14 orders of magnitude below the smallest length that can be resolved by the most powerful accelerator to date, the Large Hadron Collider. Hence the strings that particles are made of are not likely to be directly observed any time soon.

Particle interactions in string theory can be depicted as strings splitting and joining, as illustrated in Fig. 19.2. One of the major attractions of the theory is that it is free from the problem of infinities that had plagued all earlier attempts to develop a quantum theory of gravity. The problem can be traced to the point-like nature of particles. When two particles collide, their

¹String theory has a peculiar history. It was first introduced in 1970 as a theory of strong interactions. However, the theory predicted the existence of a massless boson, which had no counterpart among the strongly interacting particles. So string theory was all but discarded, only to be revived several years later, when John Schwartz and Joel Scherk realized that the problematic boson had all the properties of the graviton.

energy is concentrated at a point, so the energy density and the curvature of spacetime become infinite at the time of collision. As a result, calculations of probabilities for various particle interactions often give nonsensical infinite answers. Strings, on the other hand, have a finite size, and string theory gives reasonable, finite results for all probabilities.

19.2 Extra Dimensions

The attractive features of string theory did not come without a cost. Back in the 1970s physicists discovered that the theory suffers from peculiar mathematical flaws, called *anomalies*, that lead to violations of energy conservation and other unacceptable physical processes. They also found that the strength of anomalies depends on the number of space dimensions and that anomalies completely disappear in a 9-dimensional space. In other words, string theory is mathematically consistent only if space has six extra dimensions in addition to the familiar three.

This sounds embarrassing: why would anyone even consider a theory which is in glaring conflict with reality? Let us stop for a moment to think what having extra space dimensions would feel like. Imagine a flatland—a two-dimensional world whose inhabitants are unaware of the third dimension. A resident of this world who has access to the third dimension would then be able to perform truly magical acts. For example, she would easily escape from any jail. It would also be impossible to hide anything from this person. A locked room or a safe would look just like open rectangles from the vantage point of the third dimension. We are not aware of any such phenomena in our world, so does this mean that extra dimensions do not exist?

Not necessarily. Extra dimensions could be curled up, or, as physicists say, compactified, to a very small size. A long garden hose is a simple example of compactification: it has one large dimension along the hose and another one curled up in a small circle. When viewed from a distance, the hose looks like a one-dimensional line, but close by we can see that its surface is a two-dimensional cylinder. String theory suggests that our universe may be very similar: the compact six dimensions may be as small as the Planck length and therefore nearly impossible to detect. However, the sizes of extra dimensions and the manner in which they are compactified affect the vibrational states of the strings. And the vibrational patterns in turn determine the properties of all particles and forces. Hence, the constants of nature in our 3-dimensional world, such as particle masses and the vacuum energy density, depend on the size and shape of the hidden extra dimensions.

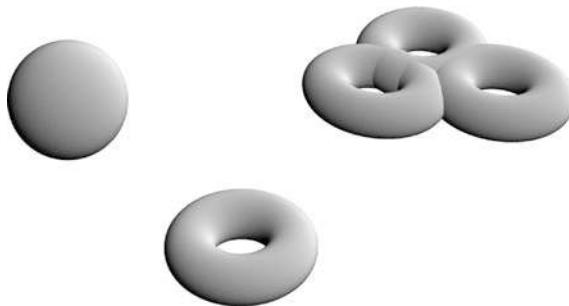


Fig. 19.3 Different ways to compactify two extra dimensions

19.3 The Energy Landscape

If we had only one extra dimension, the only way to compactify it would be to curl it up in a circle. A two-dimensional space can be compactified in a number of different ways: a sphere, a donut, or a shape with two or more “donut holes” (see Fig. 19.3). With more dimensions, the number of possibilities multiplies. Furthermore, there are other ingredients in string theory, called fluxes (these are like magnetic fields), and branes (these are membranes of various dimensionalities), which also add to the number of possible configurations that the hidden dimensions can have.

In order to fully characterize a given configuration, one has to specify the sizes and shapes of extra dimensions, the magnitudes of the fluxes that permeate them, and the locations of the branes that can wrap around them. Altogether, this amounts to specifying $N \sim 500$ different parameters. The role of these parameters in string theory is similar to that of the Higgs fields in particle physics: (i) varying the parameters results in variation of particle properties and (ii) the parameters adjust their values to minimize the potential energy density. In the simple models with one or two parameters, the energy landscape can be visually represented, as illustrated in Figs. 18.9 and 18.11. The energy minima correspond to valleys in the landscape. A similar representation for the energy landscape of string theory would require a space of N dimensions, with one dimension for each of the $N \sim 500$ parameters.

If we try to enumerate the distinct ways the extra-dimensional ingredients can be combined to form a minimum (or “valley”) in the energy landscape, we find that there are googols of possibilities. For a very rough estimate, suppose that each of the N parameters can take p different values in the valleys. The total number of possible combinations is then p^N (see Question 6). With $p \sim 10$ and $N \sim 500$, this gives 10^{500} —a truly enormous number! (by comparison, the number of atoms in the observable part of the universe is “only” $\sim 10^{80}$).

Each valley in the energy landscape corresponds to a different possible world, with its own particles, interactions and constants of nature. Thus, although strings obey a unique set of laws in higher dimensions, the many compactification choices and extra-dimensional ingredients lead to a huge ensemble of lower-dimensional vacuum states.

This raises many questions. If string theory is correct, then one out of the googols of vacuum states in the landscape corresponds to our world. But which one? How was this particular state selected to be realized in nature? And what about all the states which are not like ours? What kinds of universes do they describe? In some of them gravity may be stronger than the strong nuclear force. Others may have three different kinds of photon, and still others no photons at all. There may be states with more or less than six dimensions compactified, so the number of the remaining large spatial dimensions is different from three. Do these states exist only as possibilities, or could they exist somewhere in the physical spacetime?

19.4 String Theory Multiverse

The hope of string theorists was that the theory would yield a unique vacuum state—presumably ours. They searched for a guiding principle that would select this particular vacuum in the energy landscape. However, no plausible vacuum selection principle has yet been found. Instead, a very different picture has emerged. It was first suggested by Raphael Bousso and Joseph Polchinski, who combined string theory with the ideas of eternal inflation.

Bousso and Polchinski asserted that there are no preferred vacuum states: all vacua should be treated on an equal footing. Suppose the universe begins in a vacuum state corresponding to some valley in the landscape. If the energy density of this vacuum is positive, it will drive exponential inflationary expansion. The initial vacuum is classically stable (as are all vacuum states in the landscape), but sooner or later bubbles of other vacua will begin to nucleate by quantum tunneling through energy barriers to the neighboring valleys. Interiors of positive-energy bubbles will also be inflating and will become sites of further bubble nucleation. In this way, each type of vacuum permitted by the string theory landscape will populate the spacetime. The number of bubbles of all possible types will grow without bound during the course of eternal inflation. The resulting multiverse will look like Fig. 18.10, except that it will include $\sim 10^{500}$ different kinds of bubble universes, so it would require $\sim 10^{500}$ colors to depict it!

Most of the string theory practitioners initially viewed this multiverse idea as a giant step backwards. “This is a dangerous idea that I am simply unwilling to contemplate”, wrote the prominent Princeton cosmologist Paul Steinhardt. If the multiverse includes a multitude of different types of bubble universes, how can we ever hope to explain the observed particle properties? Whatever these properties are, we can always expect to find a suitable fit among the googols of vacua in the landscape. This looked very discouraging—a theory that can explain anything may eventually explain nothing at all.

Another approach, advocated by Bousso and Polchinski and by one of the string theory pioneers Leonard Susskind, was to embrace the string multiverse picture and explore where it leads. This approach has been steadily gaining ground among physicists in recent years. If the multiverse picture turns out to be correct, it will have far-reaching consequences for the way in which physicists go about studying the nature of the world, as we will discuss in the next chapter.

19.5 The Fate of Our Universe Revisited

In Chaps. 8 and 9 we addressed the question of the fate of our universe. The discovery of dark energy led us to conclude that the universe will continue to expand faster and faster: distant galaxies will be pushed away from each other with acceleration, but bound systems, like our Galaxy and the Local Group, will remain bound. Although the Milky Way will merge with Andromeda, most of the galaxies that we see today (except those in our Local Group), will eventually be pushed beyond our cosmic horizon. Our descendants will not see a universe filled with hundreds of billions of galaxies, as we do, but rather will find themselves in a lone island galaxy surrounded by almost nothing.

But there is more to the story. If the string landscape picture is correct, then the enormous number of vacuum states must include some positive and some negative energy vacua. This means that our vacuum does not have the lowest possible energy and must be unstable. In other words, we must be living in a false vacuum! Inevitably, a negative-energy bubble will form in our cosmic neighborhood and start to expand, engulfing more and more space. Exactly when this is going to happen is impossible to predict. Bubble nucleation can be extremely slow and can take googols of years. But on the other hand, we cannot exclude the possibility that an expanding negative-energy bubble is charging toward us at this very moment. If so, it will come without a warning: any light the bubble emits will not get to us much ahead

of the bubble itself, since it expands at nearly the speed of light. Once it arrives, our world will be completely annihilated, and all objects will be turned into some alien forms of matter.

The stage is now set for the final act of the drama. As we discussed in Chap. 18, negative vacuum energy is gravitationally attractive and will cause the interior of the bubble to contract and eventually collapse to a big crunch. This will be the end of our local region. In the meantime, outside of the crunching bubble inflation will continue and countless new bubbles will be formed. The inflating multiverse will go on forever.

Summary

String theory is perhaps the best candidate we now have for the fundamental theory of nature. It asserts that the basic building blocks of matter are one-dimensional strings. All particles of the Standard Model are thought to be tiny vibrating strings that appear point-like because the strings are so small. Different vibrational patterns give rise to distinct particles.

String theory automatically includes gravity. However, the theory is mathematically consistent only if space has 6 extra dimensions—in addition to the 3 we are familiar with. These extra dimensions are curled up, or compactified, so they are very small and we don’t notice them directly. However, the sizes of extra dimensions and the manner in which they are compactified affect the vibrational states of the strings. Hence, properties of our 3-dimensional world, such as particle masses and the vacuum energy density, depend on the size and shape of the hidden dimensions.

It turns out that there are a huge number of different ways to compactify the extra dimensions. Each one corresponds to a different possible world, or vacuum state, with its own particles, interactions and constants of nature. This ensemble of vacuum states is called the string theory landscape.

Combining string theory with the theory of inflation, we arrive at the picture of a multiverse, where bubbles of all possible vacua nucleate and expand, while inflation continues ad infinitum. If this picture is correct, then eventually an expanding bubble of negative vacuum energy will nucleate and engulf our local universe. The negative-energy bubble interior will finally collapse to a big crunch.

Questions

1. Did string theory make any predictions that have been confirmed by experiments? If not, do we have any reasons to believe that string theory is correct?
2. Why is it so difficult to test string theory observationally?

3. String theory is a candidate for a unique physical theory, from which all of physics can be derived. Does this mean that the theory should predict the observed properties of elementary particles?
4. What are some of the most surprising predictions of string theory?
5. Is it possible that there are more than three spatial dimensions in our universe? If so, why don't we see them?
6. If you have 5 different pants in your closet and five different shirts, how many distinct outfits can you make? What if you are then given five different hats—how many pant/top/hat outfits can you now make? If there are 100 extra-dimensional parameters that can each take on one of 2 values, how many possible states can be formed?
7. The string theory landscape provides a vast menu of possible types of vacua. How does the multiverse come to be populated with each and every one of these possible types?
8. What will be the ultimate fate of our observable universe, if the string landscape picture is true? How does this fate differ from the fate of our observable universe if our vacuum is completely stable?

20

Anthropic Selection

The properties of every object in the universe, from subatomic particles to giant galaxies, are determined, in the final analysis, by a set of numbers that we call “constants of nature”. These include the speed of light, Planck’s constant, and Newton’s gravitational constant; the parameters in the Standard Model, like the mass of the electron, Higgs boson, quarks and so on, and the strengths of the four forces. There are also several cosmological parameters that shape the character of our world. These include the relative contributions of radiation, atomic matter, dark matter and dark energy to the density parameter and the magnitude of the initial density inhomogeneities. Altogether there are about 30 numbers,¹ which beget an intriguing question: Why do these numbers take the particular values that they have? It has long been a dream of physicists to be able to derive all the constants of nature from some fundamental theory. But there has been very little progress in this direction.

If you write the known constants of Nature on a piece of paper, they look pretty random (see Fig. 20.1). Some of them are very small, others large, so there seems to be no system behind these numbers. However, some people noted that there may be a system, but not of the kind that physicists have been hoping for. The values of the constants appear to be fine-tuned to allow for the existence of life. In other words, if we ask what would happen if we

¹The values of the constants depend on the units we use to measure them. Physicists therefore focus on dimensionless combinations of the constants, like the ratios of particle masses, which do not depend on the units. The number we quote (30) is the number of independent dimensionless combinations of the constants.

Photon and Gluon	0
W-boson	157 000
Electron	1
Neutrino	$< 10^{-8}$
Muon	207
Up-quark	8
Bottom quark	9200

Fig. 20.1 Masses of some particles, in units of the electron mass. The values appear to be rather random

change one of the constants by a relatively small amount, we find that we would get a universe that is inhospitable to life. Let us consider a few examples that illustrate how tinkering with the constants leads to catastrophic results.

20.1 The Fine Tuning of the Constants of Nature

20.1.1 Neutron Mass

The mass of the neutron is very finely tuned. If we adjust it just a little, we change the structure of matter so much that chemistry is almost completely destroyed. Let's see why. Neutrons are 0.14% heavier than protons. Outside of the nucleus, they decay into protons, electrons and antineutrinos: $n \rightarrow p^+ + e^- + \bar{\nu}$. These "free" neutrons have an average life of about 15 min. But inside nuclei neutrons are stabilized by nuclear forces. If the neutron's mass were *increased* by 1%, then neutrons would decay even inside nuclei, turning into protons. The electric repulsion between the protons would then tear the nuclei apart, so the only stable nucleus would be that of hydrogen, consisting of a single proton. On the other hand, if the neutron's mass were *decreased* by 1%, neutrons would become lighter than protons. This would mean that protons would decay into neutrons, positrons and neutrinos: $p^+ \rightarrow n + e^+ + \nu$. Consequently, atomic nuclei would lose their charge and would consist only of neutrons. The unattached electrons would

fly away, so no atoms would exist. Thus, by adjusting the mass of the neutron just a little, we either end up in a world that only contains one type of chemical element—hydrogen—or a neutron world.²

20.1.2 Strength of the Weak Interaction

When a massive star runs out of nuclear fuel, its core collapses in a supernova explosion. The strength of the weak interaction is perfectly suited to allow neutrinos to stream out of the core and drag along the outer layers of the star. This is a critical part of the cycle that enriches the interstellar medium with heavy elements. If weak interactions were much stronger, neutrinos would remain stuck in the core. If they were much weaker, neutrinos would stream out without dragging along other particles. If the heavy elements were not spewed into space, later generations of stars and planetary systems like ours would not have formed, and the raw materials for complex life would be missing.

20.1.3 Strength of Gravity

Gravity is by far the weakest force—it is 10^{36} times weaker than electromagnetism. Because gravity is so weak, we can increase its strength quite a lot, and it will still be weak. For example, if we make it ten billion times stronger, it would still be 10^{26} times weaker than electromagnetism. Stars would then be the size of mountains, and they would live for only a year or so. Intelligent life would hardly have enough time to evolve. Planets as massive as the Earth would be about 100 m in diameter, and the force of gravity on their surface would crush any object heavier than an ant.

20.1.4 The Magnitude of Density Perturbations

Structure formation in the universe crucially depends on the magnitude of primordial density perturbations. If these perturbations were much weaker, then galaxies may never have coalesced. (Note that structure formation

²On a more fundamental level, protons and neutrons are made up of quarks, so it is more appropriate to regard the quark masses as fundamental constants of nature. But the general conclusion does not change: we are driven to either a hydrogen world or a neutron world, unless the quark masses are suitably fine-tuned.

freezes at the onset of dark energy domination, thus if galaxies fail to form prior to this epoch, they will never form.) Without galaxies there would be no buildup of heavy elements, and it is unlikely that planets, and life, would have emerged.

If the initial density perturbations were much stronger, then galaxies would form earlier and would be much denser. Close stellar encounters would be much more frequent; they would disrupt planetary orbits, with disastrous consequences for life.

20.2 The Cosmological Constant Problem

We now come to the most striking fine-tuning of all. The observed accelerated expansion of the universe is caused by a vacuum energy (mass) density, or cosmological constant, which is about twice the average density of matter today, $\rho_v \sim 2\rho_m$. This value is in blatant conflict with theoretical expectations.

20.2.1 The Dynamic Quantum Vacuum

When you think of a vacuum, you intuitively picture a state of pure “emptiness” or “nothingness”. However, quantum theory tells us that the vacuum is an inextinguishable sea of virtual particles that spontaneously appear and disappear. All particles in the Standard Model—electrons, quarks, photons, W-bosons, and so on—are relentlessly fluctuating in and out of existence. Although these virtual particles are very short lived, they have important and measurable effects.³ Most importantly, they contribute to the energy density of the vacuum. The problem is, however, that calculations of the resulting vacuum energy density give values that are absurdly large, $\rho_v \sim 10^{120} \rho_m$ (see the box at the end of this section). So we seem to have a mismatch between theory and observation that is about 120 orders of magnitude! This has been called “the worst prediction in physics”, “the mother of all physics problems”, or less dramatically, the “cosmological constant problem”.

³One of these is the Casimir effect which predicts that there will be an attractive force between two uncharged parallel conducting plates in a vacuum. The reason is that electromagnetic field fluctuations are restricted between the plates and unrestricted outside them. This results in more pressure from the outside pushing the plates towards one another. This effect has been measured. Also, the energy levels of the hydrogen atom have been measured and agree with the theory to a very high precision if we take into account the virtual particles which swarm inside the hydrogen atom.

Why is the observed value of the vacuum energy density so small? Is it possible that some mechanism could cause contributions from different particle species to cancel one another? It turns out that fermions and bosons do indeed contribute to the vacuum energy density with opposite signs. Bosons have a positive contribution and fermions contribute a negative energy density.⁴ But these contributions would need to cancel precisely to the 120th decimal point in order to predict a value that is as low as measured by the supernovae observations. Such a precise cancellation would be a dramatic example of fine-tuning.

20.2.2 Fine-Tuned for Life?

Let us now see what would happen if the value of the cosmological constant were very different from what it actually is. Suppose first that ρ_v is positive and is 1000 times greater than ρ_m . It would still be 117 orders of magnitude below its theoretically expected value.

The vacuum energy would then start dominating the universe at $t \sim 0.5$ Byr. At that time, galaxy formation was just beginning and only very small galaxies had enough time to form. But once the vacuum energy dominates, galaxy formation comes to a halt. The problem with miniature galaxies is that their gravity is too weak to keep heavy elements expelled in supernova explosions from flying away into outer space. Thus the galaxies would be left without the elements necessary for the formation of planets and for the evolution of life. If we further increase ρ_v by another factor of 100, then it would come to dominate well before the epoch of galaxy formation, and the universe would be left with no galaxies at all.

Suppose now that ρ_v is negative and has magnitude 1000 times greater than ρ_m . Then the gravity of the vacuum would be attractive and would cause the universe to contract and collapse to a big crunch at $t \sim 0.5$ Byr. This is hardly enough time for the evolution of intelligent life (which took about 10 times longer here on Earth). A further increase in the magnitude of ρ_v would make the lifetime of the universe even shorter and the evolution of life and intelligence even less likely.

⁴The reason for this difference is that fermions are mathematically described by so-called Grassmann numbers, which are rather different from ordinary numbers. When you multiply ordinary numbers, the result does not depend on the ordering of the factors; for example, $3 \times 5 = 5 \times 3$. But for Grassmann numbers the product changes sign under factor ordering: $a \times b = -b \times a$.

Virtual particles and vacuum energy density

According to quantum physics, the vacuum is awash with virtual particles constantly popping in and out of existence. Particles and antiparticles appear in pairs and almost instantly annihilate. The lifetime of a virtual pair Δt depends on the energy of the particles E : the higher the energy, the shorter is the lifetime. Quantitatively, this can be expressed as $E \cdot \Delta t \sim \hbar$, where \hbar is the Planck constant. Virtual particles move at nearly the speed of light, so the whole process occurs on a length scale $L \sim c\Delta t \sim \hbar c/E$. Space is packed with virtual pairs, and once a pair annihilates, another instantly appears in its place. So, if you look at a small cubic region of size L at any time, you are likely to find a pair of particles with energies $E \sim \hbar c/L$.

The energy density due to the virtual particles can now be estimated by dividing the energy E by the volume L^3 :

$$E/L^3 \sim \hbar c/L^4 \quad (20.1)$$

As L is decreased, the energy density grows, indicating that energetic pairs popping out on smaller distance scales give a greater contribution to the vacuum energy density. As we include virtual pairs on smaller and smaller scales, the energy density appears to grow without bound.

However, there may be a limit to how small the length L can be. At super-small distances, quantum gravity effects become significant and the geometry of spacetime undergoes large quantum fluctuations. Below a certain characteristic distance, spacetime acquires a chaotic, foam-like structure (see Fig. 20.2). We can estimate this distance scale using dimensional analysis. It can only depend on the fundamental constants \hbar , c and G , and the only combination of these constants that has the dimension of length is

$$\ell_p = \sqrt{\frac{\hbar G}{c^3}} \quad (20.2)$$

This is the Planck length, which we introduced in Sect. 19.1. On much larger scales, the spacetime appears to be smooth, just as the foamy surface of the ocean appears smooth when viewed from an airplane.

The physics of spacetime foam is not well understood, but physicists expect the virtual pair production to cease on scales smaller than ℓ_p . (This fits well with string theory, where the typical size of vibrating strings is $\sim \ell_p$.) The vacuum energy density can then be estimated by setting $L \sim \ell_p$ in Eq. (20.1) and the corresponding mass density can be obtained by further dividing by c^2 :

$$\rho_v \sim \frac{c^5}{\hbar G^2} \sim 10^{97} \text{ kg/m}^3 \quad (20.3)$$

This is greater than the observed vacuum energy density by a factor of about 10^{123} .

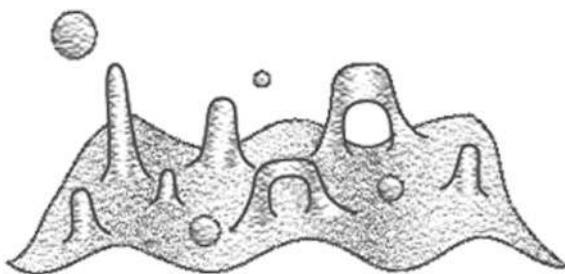


Fig. 20.2 As space is viewed at higher and higher resolution, a foamy structure emerges at the Planck scale

20.3 The Anthropic Principle

Why are the constants of nature fine-tuned for life? There are a few ways to address this question, and we shall consider them in turn.

1. The universe is what it is. The constants have to have some values, and they just happen to be consistent with life. We simply got lucky. This is not very satisfactory: getting lucky many times in a row calls for an explanation. And a fine-tuning by 120 orders of magnitude is hard to dismiss as simply an accident.
2. One day we will have a complete theory of physics that will allow us to calculate all the parameters from first principles. We just have to buckle down and keep working towards such an understanding. But how likely are the constants derived from the fundamental theory to fall in the narrow ranges allowing life to exist? If they do, that would be a tremendous stroke of luck. Once again, it would not be satisfactory to leave it unexplained.
3. The constants were fine-tuned by a benevolent creator, just so we can exist. There is often a temptation to invoke God whenever we encounter something that seems very hard to explain. But this “God of the gaps” approach has a poor success record in science. Isaac Newton, for instance, suggested that a supernatural deity was responsible for sustaining a homogeneous distribution of stars against gravitational collapse and for the fact that the planets are “opaque” and the stars “luminous”. Of course, it has been a great triumph of science to discover the expansion of the universe and to explain how thermonuclear reactions cause an opaque body to become a luminous star.

- Finally, there is a possibility that the constants of nature can take on a variety of different values, which can be realized in distant parts of the universe beyond our horizon. Then we should not be surprised to find ourselves living in a *special* corner of the universe which has constants of nature that are hospitable to life. We cannot live in environments that are not bio-friendly—even if most of the universe is of this sort. We live only where we can! This is the so-called “anthropic principle”.

To illustrate the anthropic principle at work, let us think about the Solar System for a moment. About four centuries ago, when the Solar System was thought to be the universe, Johannes Kepler asked the following question: What determines the number of planets and their particular distances from the Sun? At that time only five planets were known, and Kepler was struck by the fact that this was exactly the number of highly symmetric polyhedrons, called Platonic solids.⁵ He came up with an elaborate construction where the solids were nested inside one another and suggested that their sizes were proportional to the radii of planetary orbits (see Fig. 20.3). But today it is obvious that Kepler was asking the wrong question. We have detected thousands of extrasolar planets, and we have every reason to expect that there are billions of them in the observable universe. The planets orbit their suns at a great variety of distances, but most of them are not well suited for the evolution of life. If our Earth were significantly closer or farther away from the Sun, the oceans would either boil or freeze, and life of our kind would be impossible. The reason why we live on a planet that is “hospitably” located is simply because we can’t live on a planet at an inhospitable distance. If the Solar System were the only one in the universe, then it would be very mysterious that it contains a bio friendly planet. But if there are many types of planets with varied conditions, then it is not so surprising that some of them have environmental factors that are hospitable to life—and it is common sense that we live on such a planet.

Similarly, if we are living in a multiverse, where there are many distant regions that have different constants of nature, then it is not at all surprising that we find ourselves in a very special place with “fine-tuned” parameter values. We simply can’t live anywhere else. In the multiverse context, asking why a given parameter has a specific value is to ask the wrong question—like Kepler.

⁵This fact was discovered by the ancient Greeks. The Platonic solids are the tetrahedron, cube, octahedron, dodecahedron and icosahedron.

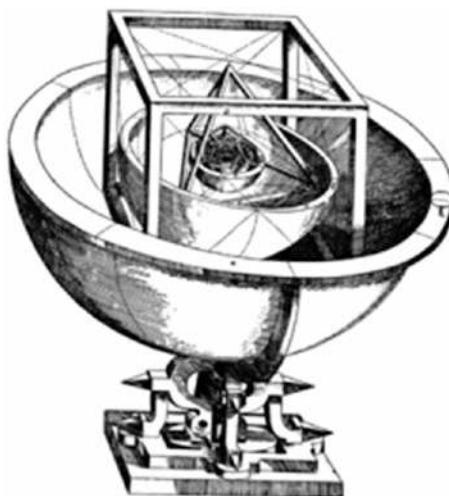


Fig. 20.3 Kepler's model of the Solar System with five Platonic solids nested within one another

The anthropic principle, which was introduced in 1974 by the Australian born astrophysicist Brandon Carter, has a dubious reputation among physicists. On the one hand, the principle is trivially true (we cannot live where life is impossible), and its application to our location in the Solar System is uncontroversial. On the other hand, its use for explaining the fine-tuning of the constants of nature has often been viewed with great suspicion.

20.4 Pros and Cons of Anthropic Explanations

Anthropic explanations assume the existence of a multiverse consisting of remote domains where the constants of nature take different values. In the 1970s this assumption appeared to be rather far-fetched, but this has now changed due to subsequent developments in particle physics and cosmology. Modern particle theories predict the existence of multiple vacuum states with diverse properties, and eternal inflation provides a mechanism for populating the universe with large regions of all possible vacua.

Furthermore, to explain the fine-tuning of the vacuum energy density ρ_v , the number of vacua in the energy landscape of the theory should be enormous. To understand why, let us imagine a long ribbon representing possible values of ρ_v , from $-10^{120} \rho_m$ to $+10^{120} \rho_m$. At the center of the ribbon is a minuscule anthropic range, between $-10^3 \rho_m$ and $+10^3 \rho_m$, where life is possible. Now, we want the number of vacua in the landscape to be sufficiently large, so that some of them happen to be located in the anthropic range. If we randomly throw a dart at the ribbon, the probability that it will hit the anthropic range is completely negligible,

$$P \sim \frac{10^3 \rho_m}{10^{120} \rho_m} \sim 10^{-117}. \quad (20.4)$$

We will have to make more than 10^{117} attempts before we can expect to have a successful hit. Similarly, we need an energy landscape of more than 10^{117} vacua for the anthropic explanation of ρ_v to be successful.

Energy landscapes of grand unified theories typically include only a few vacua and fall far short of the mark, and this is where string theory comes to the rescue. As we discussed in Chap. 19, the energy landscape of string theory is estimated to have $\sim 10^{500}$ vacua. This completely dwarfs the required number 10^{117} . With such an immense landscape, we can expect to have googols of vacua in the anthropic range (see Question 7).

The anthropic principle has often been dismissed as being unpredictive and untestable—a philosophical cop out. It gives a ready explanation for any values of the constants of nature that we can measure, but does not seem to provide any means to verify that this explanation is correct. Today, however, many physicists are realizing that anthropic arguments may in fact lead to testable predictions, as we shall discuss in the next chapter.

Summary

In our observable universe there are roughly 30 constants of nature that have been measured empirically. Despite their best efforts, physicists have not been able to derive the values of these parameters from first principles. Interestingly, a relatively small change to the value of any of the constants tends to lead to a universe that is inhospitable to life. How can we explain this fine-tuning?

One possibility is that the constants of nature can take on a variety of different values, which can be realized in distant parts of the universe beyond our horizon. Then it is no surprise that we live in a fertile zone that has constants of nature that are hospitable to life. We cannot live in environments

that are not bio-friendly—even if most of the universe is of this sort. This is the so-called “anthropic principle”.

The observed vacuum energy density, or cosmological constant, is about 120 orders of magnitude smaller than the theoretical value. This is called the “cosmological constant problem”, and it is one of the biggest mysteries in theoretical physics. The anthropic principle, combined with the multiverse worldview, can be used to explain why the cosmological constant is so small.

Questions

1. Neutrons are slightly heavier than protons. What would happen if we could decrease the neutron mass by 1%, so that protons become the heavier of the two?
2. The orbit of the Earth around the Sun is nearly circular, while many of the extra solar planets are observed to have highly eccentric elliptical orbits. Why do you think we do not live on one of those planets?
3. Give two examples of constants of nature which appear to be fine-tuned. For each example indicate one way in which the universe would be very different if these constants had different values.
4. What is the “anthropic principle”?
5. Explain why a high value of vacuum energy density hinders galaxy formation.
6. How does the idea of a multiverse explain the apparent fine-tuning of the cosmological constant?
7. Using the expression for the typical energy of the virtual pairs in the box at the end of Sect. 20.2, find the length scale L below which the particles of the pair would form a black hole. This is one of the ways to find the length scale at which quantum gravity effects become important. Does your answer agree with the result of the dimensional analysis in the box? (Hint: Particles having combined mass M form a black hole if they are localized within a sphere of radius smaller than the Schwarzschild radius $2GM/c^2$.)
8. Suppose the range of possible values of ρ_v is from $-10^{120}\rho_m$ to $+10^{120}\rho_m$, and the anthropic range allowing for the existence of life is from $-10^3\rho_m$ to $+10^3\rho_m$. Furthermore, suppose the energy landscape includes 10^{500} vacua. Estimate the number of vacua in the anthropic range.

21

The Principle of Mediocrity

According to the multiverse worldview, the constants of nature vary from one part of the universe to another. In some regions the constants allow for the existence of life, and that is where observers will evolve and those constants will be measured. Observers in different regions will generally measure different values of the constants. We don't know *a priori* what kind of region we live in, so we cannot predict the local values of the constants with certainty. However, it may be possible to make *statistical* predictions. Some region types may be more numerous or more densely populated than others, and we are more likely to find ourselves in one of these more populous regions.

21.1 The Bell Curve

If you have ever taken a large introductory college course, you have probably wondered if you are being graded on a “curve”. The curve of course, is the so-called “bell curve” (see Fig. 21.1). What does the bell curve represent? Let’s suppose the same final exam is given to a class of 300 students, every year for 20 years. If you were to randomly pick a name from a hat (containing the names of all students who have taken the class), what grade do you expect that student to have attained on their final exam? You would be surprised if the student’s grade were, say, in the top or bottom 1% of the class. If the teacher supplied you with a set of data plotting the results of students who took that final for the last 20 years, as shown in Fig. 21.1, you

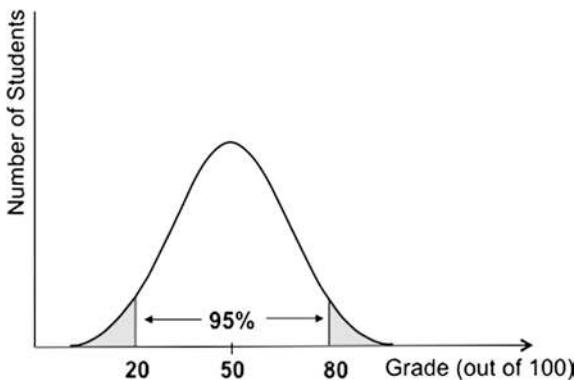


Fig. 21.1 Grade distribution of students in a large class. The number of students whose grades are within a specific range is proportional to the area under the corresponding piece of the curve. The median grade is 50 points. This means that half of the students got grades above, and the other half below, this value. The shaded tails represent the lowest and highest 2.5%. The range of grades between the two shaded areas is predicted at 95% confidence level

would be able to make more accurate predictions. If you discard the 2.5% of students who got the highest and lowest grades, then with 95% confidence you can say that a randomly picked student got a grade in the remaining 95% range (that is, between 20 and 80 out of 100 in the example shown in Fig. 21.1). This means that if you were to pick name after name, then 95% of the time you would pick students who scored in the aforementioned range. This is called a prediction at 95% confidence level. In order to make a 99% confidence prediction, you would have to discard 0.5% at both ends of the distribution. As the confidence level is increased, your chances of being wrong get smaller, but the predicted range of grades gets wider and the prediction less interesting.

21.2 The Principle of Mediocrity

A similar technique can be applied to make predictions for the constants of nature. Suppose for a moment that there is a Universal Super Observer who can survey all of spacetime and measure anything she wants. The USO looks around and sees many different regions of the universe with different values for the constants of nature. She decides to count the number of observers who live in regions that have different values of a certain constant, call it X. To isolate the effect of varying X, she focuses on regions where the other

constants have nearly the same values and only X changes from one region to another.¹ The USO finds that some of the regions contain many observers, some only a few, and others have none. She can then plot the number of observers who will measure various values of X. The resulting distribution will most likely be similar to a bell curve. If the USO graciously gives us the distribution, we could discard 2.5% at both of its ends and make a 95% confidence level prediction for the value of X measured by a randomly selected observer.

What would be the use of such a prediction? Obviously, we would not be able to test it directly—we can't “pick up the phone” and ask randomly selected observers to disclose their measurements—because all regions with different values of X are beyond our horizon. What we can do, though, is to think of ourselves as having been randomly selected. Since we have no a priori reason to believe that the values of the constants in our region are unusually large or small, or otherwise very special, it makes sense to assume that we are typical, or unexceptional observers. This assumption is called the *principle of mediocrity*. If there are some constants of nature that we have not yet measured, and if we have somehow obtained the statistical distribution for their values measured by all the observers in the multiverse, we can use the principle of mediocrity to predict that the values of the constants in our local region should correspond to the range around the peak of the distribution.

But where are we going to get the distribution? In lieu of a cooperative USO, we will have to derive it from our theory of the multiverse. If the resulting predictions agree with our measurements, this would provide evidence supporting the theory; if not, the theory can be ruled out at a specified confidence level.

21.3 Obtaining the Distribution by Counting Observers

Let us now discuss how the distribution of the constants measured by randomly picked observers can be derived from the theory. We have to count the number of observers in regions with different constants. In order to do

¹If the vacuum landscape is indeed as rich as string theory suggests, it will include vacua with practically any values of the constants. So the USO will have no problem finding regions where X varies while the other constants are nearly fixed.

so, we need to know the density of observers in each environment, and the corresponding volume. The volume factor can in principle be calculated from the theory of inflation (we'll discuss this in Sect. 21.5). However, since we are mostly ignorant about the evolution of life and intelligence, how can we hope to calculate the density of observers that may arise in different environments?

For starters, we should note that some constants of nature are “life-changing” and some are “life-neutral”. A “life-changing” constant is one that directly affects the physics and chemistry of life, and thus directly impinges on the ability of life to evolve and thrive within a galaxy. Examples of life-changing constants include the electron mass and the gravitational constant. On the other hand, a “life-neutral” constant is one that, as long as there are galaxies around, does not directly influence the ability of life to emerge. For example, the cosmological constant and the magnitude of the primordial density fluctuations are life-neutral constants. As long as their values lie within the windows that permit the formation of galaxies in a region, they do not influence the density of observers within any galaxy.²

At the present level of understanding, we can only attempt to calculate the distributions of life-neutral constants. Furthermore, our ignorance about the emergence of life can be factored out if we focus on those regions where the life-changing constants have the same values as in our neighborhood, and only the life-neutral constants are different. All galaxies in such regions will have about the same number of observers, and thus to compare the density of observers in different regions, we only need to compare the density of galaxies. In other words, *we can use the density of galaxies as a proxy for the density of observers*. We shall now discuss how this approach was used to predict the value of the cosmological constant.

21.4 Predicting the Cosmological Constant

The anthropic bound on the cosmological constant ρ_v specifies the value above which the vacuum energy would dominate too soon for any viable galaxies to form. In regions where ρ_v is near this bound, galaxy formation is barely possible, and the density of galaxies is very low. But most observers

²This is a bit of an oversimplification. Some properties of galaxies may in fact change due to variation of life-neutral constants. For example, if the density fluctuations get larger, galaxies form earlier and have a higher density of matter. As a result, close encounters between stars, which can disrupt planetary orbits and extinguish life, become more common.

will not live in these lonely places; they will live in regions that are teeming with galaxies. Thus, if we assume that we are typical observers, we should expect to live in one of the galaxy-rich regions and to measure a cosmological constant that is significantly lower than the anthropic bound.

21.4.1 Rough Estimate

A rough estimate of the expected value of ρ_v can be obtained as follows. Let us consider a large ensemble of regions where ρ_v takes a variety of values, while the other constants are very close to what they are in our local neighborhood. Depending on the value of ρ_v , the vacuum energy in these regions will start dominating at different times t_v , or different redshifts z_v . Once the vacuum energy dominates, galaxy formation comes to a halt, so regions where the vacuum domination occurs after only a few galaxies have had a chance to form, will have sparse observers.

As we discussed in Chap. 12, galaxy formation proceeds in a hierarchical manner, with smaller clumps merging to form larger and larger structures. Large galaxies like ours, massive enough to efficiently form stars and to retain the heavy elements dispersed in supernova explosions, are formed at redshifts $z \sim 2$ or later. In galaxy-rich regions the vacuum domination should occur at a later time, and thus we must have $z_v < 2$ (Remember: smaller redshifts correspond to later times.). Now, the density of matter at $z = 2$ is $\rho_m = (1 + z)^3 \rho_{m0} = 27 \rho_{m0}$, where ρ_{m0} is the present value of ρ_m .

Requiring that the vacuum energy does not dominate at this epoch, we obtain (see Sect. 5.1 in Chap. 5)

$$\rho_v < \frac{\rho_m}{2} \approx 14\rho_{m0} \quad (21.1)$$

The values of ρ_v measured by most observers in the multiverse living in environments similar to ours are expected to satisfy this condition.

21.4.2 The Distribution

The probability distribution for the values of ρ_v measured by randomly picked observers requires a more careful calculation. The result of such a calculation is plotted in Fig. 21.2. The distribution is peaked at $\rho_v \sim 3\rho_{v0}$, where $\rho_{v0} \sim 2\rho_{m0}$ is the observed value, and the 95% confidence range is between $\sim 0.1\rho_{v0}$ and $\sim 20\rho_{v0}$. Values of $\rho_v > 20\rho_{v0}$ are not likely to be observed because there are very few galaxies in the corresponding regions. Very small

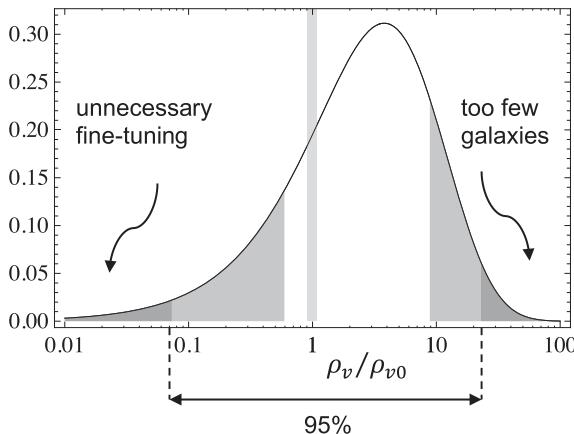


Fig. 21.2 Probability for a randomly picked observer to measure a given value of ρ_v . The *dark grey* and *light grey areas* mark the values excluded at 95 and 67% confidence levels, respectively. The *vertical bar* marks the observed value. From A. De Simone, A. Guth, M. Salem and A. Vilenkin, Phys. Rev. D78, 063520 (2008)

values of $\rho_v < 0.1\rho_{v0}$ are also unlikely, simply because this range of values is so narrow. A value in this range would amount to unnecessary fine-tuning; that is, a fine-tuning even more severe than required by the anthropic considerations.

Throughout this section, we implicitly assumed that $\rho_v > 0$. A similar analysis can be performed for negative values of ρ_v , with very similar conclusions. The main difference is that the bound on large negative values of $\rho_v < -20\rho_{v0}$ comes from requiring that the universe does not collapse to a big crunch before some galaxies manage to form.

Anthropic bounds on ρ_v were first derived in 1987 by Steven Weinberg and by Andrei Linde. A prediction based on the principle of mediocrity was made by Vilenkin in 1995 and was later refined by George Efstathiou (1995) and by Hugo Martel, Paul Shapiro and Weinberg (1998). At the time anthropic arguments were highly unpopular,³ and it came as a complete surprise to most physicists when a vacuum energy density of roughly

³The referee of the Astrophysical Journal objected to publishing papers based on anthropic reasoning, so in order for Martel Shapiro and Weinberg to get their 1998 paper accepted, they had to convince the editor, that if ρ_v was ever measured to be below a certain value, this would show that anthropic reasoning could not explain it. Of course, the value of ρ_v turned out to be just in the sweet spot for an anthropic explanation to make perfect sense.

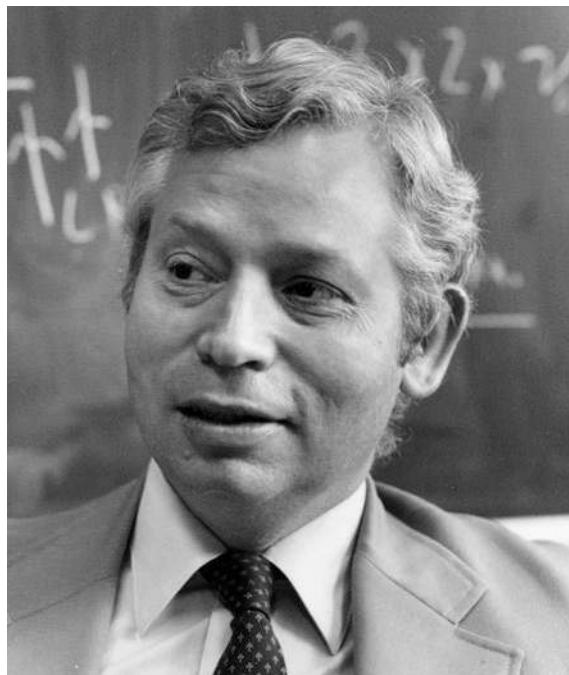


Fig. 21.3 Steven Weinberg won the 1979 Nobel Prize for his work on the Standard model of particle physics. He also made seminal contributions to cosmology. Credit AIP Emilio Segré Visual Archives

the expected magnitude was detected in supernova observations in 1998. As of now, no viable alternative explanations for the observed value of ρ_v have been proposed. This may be our first observational evidence for the existence of a multiverse (Fig. 21.3).

21.5 The Measure Problem

In order to calculate the probability distribution for values of a certain constant measured by randomly picked observers, we need to know the fraction⁴ of volume of the universe where various values of the constants are realized, and also the density of observers in each of these environments. For life-neutral constants the density of observers is proportional to the density

⁴See question 4 to convince yourself that it is sufficient to use volume fractions, instead of actual volumes, to calculate probabilities.

of galaxies, which can be calculated in a relatively straightforward way. However, the calculation of the volume fraction presents a serious problem.

The problem arises because the volumes of all kinds of environments in the multiverse grow without bound and become infinite in the limit. So, to find the fraction of the volume occupied by a given environment we have to compare infinities, and this is a mathematically ambiguous task. This can be illustrated by a simple example of an infinite sequence of integers:

$$1, 2, 3, 4, 5, 6, 7, 8, \dots$$

We can ask: What fraction of the integers are odd? You probably guessed $\frac{1}{2}$. Indeed, if we take N numbers in a row, the fraction of odd numbers will be close to $\frac{1}{2}$ for large N and will exactly equal $\frac{1}{2}$ in the limit of infinite N .

But if you reorder the sequence so that each odd integer is followed by two even integers,

$$1, 2, 4, 3, 6, 8, 5, \dots$$

then the answer would be $\frac{1}{3}$ (even though this sequence contains all the same integers as the natural ordering). In fact, by reordering the sequence one can obtain any answer to this question between 0 and 1.

In this particular example the ambiguity can be avoided by requiring that the natural order of integers should be used. The answer is then $\frac{1}{2}$, as one intuitively expects. We could try to adopt a similar prescription for the volume fraction in the multiverse, using the natural ordering of events in time. This would amount to including only the regions (e.g., bubble universes) that were formed prior to a certain time t . The problem is, however, that the result depends on how time is defined in different places. There is no unique, or preferred way to do so in general relativity. One can use, for example, the time measured by the clocks of local observers; this is called the *proper time measure*. Alternatively, one could use the expansion of the universe as a measure of time. Equal times would then correspond to equal values of the scale factor; this is referred to as the *scale factor measure*. There is an infinity of possible choices, and thus the volume fraction remains ambiguous. This ambiguity is known as *the measure problem*.

Cosmologists have studied different measure prescriptions and found that some of them lead to paradoxes or to a conflict with the data and should therefore be discarded. For instance, the proper time measure performed rather poorly, while the scale factor measure has successfully passed all tests

so far.⁵ It is unlikely, however, that this kind of analysis will yield a unique prescription for the probabilities. This suggests that some important element may be missing in our understanding of cosmic inflation.

Some people feel the measure problem is so grave that it puts the validity of the theory of inflation seriously in doubt. But this is the view of only a small minority of cosmologists. The situation with the theory of inflation is similar to that with Darwin's theory of evolution some 100 years ago. Both theories greatly expanded the range of scientific inquiry, proposing an explanation for something that was previously believed impossible to explain. In both cases, the explanation was compelling, and no viable alternatives have been suggested. Darwin's theory was widely accepted, even though some important aspects remained unclear before the discovery of the genetic code. The theory of inflation may be similarly incomplete and may require additional new ideas. But it also has a similar air of inevitability.

21.6 The Doomsday Argument and the Future of Our Civilization

The principle of mediocrity has been used in many different contexts. As an example, suppose you are presented with a bag containing N cards. You know the cards are labeled 1 through N , but you don't know what N is. Now you draw one card at random and see number 15 written on it. Based on this, how would you estimate the total number of cards N ?

The principle of mediocrity suggests that your card is not likely to be from the very beginning or the very end of the list and comes most likely from somewhere in the middle. Then your best estimate is $N = 30$. If you want to make a 90% confidence prediction, this would be $16 < N < 300$. (Can you figure out how we obtained these numbers?). You can make a more accurate prediction if you draw more than one card. The Allied forces used a similar method during World War II to estimate the total number of German tanks based on the serial numbers of the tanks that they captured.

If we imagine giving a "serial number" to every person at birth, we can use the same reasoning to predict the total number of humans who will ever live. The number of people who have lived on Earth since the origin of our

⁵The distribution for the cosmological constant in Fig. 21.2 was calculated using the scale factor measure. In fact, analysis shows that this distribution is not very sensitive to the choice of measure, so the prediction for the cosmological constant is rather robust and is not expected to change much when the measure problem is finally resolved.

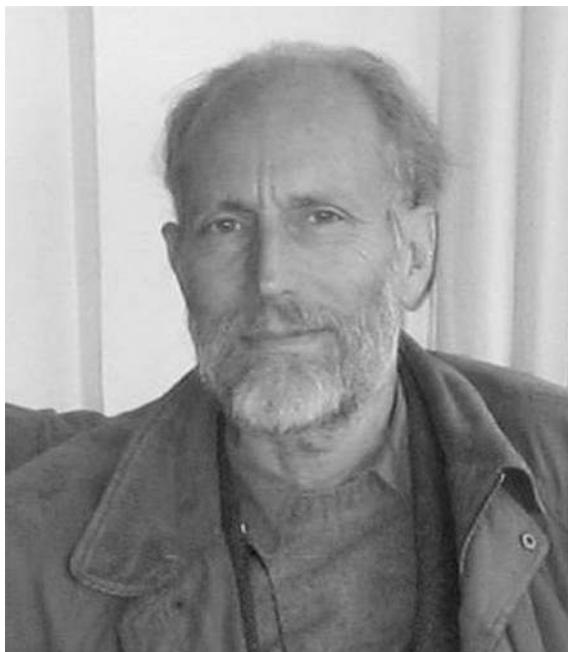


Fig. 21.4 Brandon Carter is known for his important work on the properties of black holes. He also introduced the anthropic principle and the doomsday argument. *Credit* Courtesy Brandon Carter

species is about 100 billion, so our best estimate is that 200 billion people will ever live. If the world birth rate stabilizes at its current value (130 million births per year), this number will be reached in less than 800 years.

This is the notorious “doomsday argument”, first presented by Brandon Carter in 1983. The argument becomes more subtle and the prediction less gloomy if one takes into account the existence of multiple other civilizations in the universe. It should be noted that the doomsday argument is rather controversial and many people believe that the principle of mediocrity should not be used in this context (Fig. 21.4).

21.6.1 Large and Small Civilizations

For any civilization confined to a single planet, the prospects of long-term survival are rather bleak. It can be destroyed by an asteroid impact or a nearby supernova explosion, or it can self-destruct in a nuclear war. It is not a matter of *if* but rather of *when* the disaster will strike, and the only sure

way for the civilization to survive in the long run is to spread beyond its native planet and colonize space.

An advanced civilization may colonize nearby planetary systems. The colonies may then spread further, until the entire galaxy is colonized. It is conceivable that civilizations could expand even beyond their native galaxies.

The probability for a civilization to survive the existential challenges and colonize its galaxy may be small, but it is non-zero, and in a vast universe such civilizations should certainly exist. We shall call them *large* civilizations. There will also be *small* civilizations which die out before they spread much beyond their native planets.

For the sake of argument, let us assume that small civilizations do not grow much larger than ours and die soon after they reach their maximum size. The total number of individuals who lived in such a civilization throughout its entire history is then comparable to the number of people who ever lived on Earth, about 100 billion.

A large civilization contains a much greater number of individuals. A galaxy like ours has about 100 billion stars and about 20% of the stars have habitable planets (see Sect. 13.5). This amounts to 20 billion habitable planets per galaxy. Assuming that each planet will reach a population similar to that of the Earth, we get $\sim 2 \times 10^{21}$ individuals. The numbers can be much higher if the civilization spreads well beyond its galaxy. The crucial question is: what is the probability P for a civilization to become large?

It takes more than 10^{10} small civilizations to provide the same number of individuals as a single large civilization. Thus, unless P is extremely small (less than 10^{-10}), individuals live predominantly in large civilizations. That's where we should expect to find ourselves if we are typical inhabitants of the universe. Furthermore, a typical member of a large civilization should expect to live at a time when the civilization is close to its maximum size, since that is when most of its inhabitants are going to live.

These expectations are in glaring conflict with what we actually observe: we either live in a small civilization or at the very beginning of a large civilization. With the assumption that P is not very small, both of these options are very unlikely—which indicates that the assumption is probably wrong. If indeed we are typical observers in the universe, then we have to conclude that the probability P for a civilization to survive long enough to become large must be very tiny. In our example, it cannot be much more than 10^{-10} .

21.6.2 Beating the Odds

The Doomsday argument is statistical in nature. It does not predict anything about our civilization in particular. All it says is that the odds for any given

civilization to grow large are very low. At the same time, some rare civilizations do beat the odds.

What distinguishes these exceptional civilizations? Apart from pure luck, civilizations that dedicate a substantial part of their resources to space colonization, start the colonization process early, and do not stop, stand a better chance of long-term survival. With many other diverse and pressing needs, this strategy may be difficult to implement, but this may be one of the reasons why large civilizations are so rare.

One question that needs to be addressed is: why is our Galaxy not yet colonized? There are stars in the Galaxy that are billions of years older than our Sun, and it may take less than a million years for an advanced civilization to colonize the entire Galaxy. So, we are faced with Enrico Fermi's famous question: Where are they? A possible answer is that we may be the only intelligent civilization in our Galaxy and maybe even in the entire observable universe. Evolution of life and intelligence may require some extremely improbable events, as we discussed in Chap. 13. Their probability may be so low that the nearest planet with intelligent life could be far beyond our horizon.

Summary

If the constants of nature vary from one part of the universe to another, their local values cannot be predicted with certainty, but we can still make *statistical* predictions.

We have no a priori reason to believe that the values of the constants in our region are particularly special, so it makes sense to assume that we are typical, or unexceptional observers. This assumption is called the *principle of mediocrity*. It suggests that the probability for us to measure certain values of the constants in our local region is the same as for a randomly picked observer in the multiverse. Even though our understanding of the multiverse is rather incomplete, in some special cases we can calculate probabilities from the theory of eternal inflation.

This strategy was applied to the cosmological constant and led to a prediction, which was later confirmed by the 1998 Supernova observations. This could be our first evidence for the existence of the multiverse.

In general, making statistical predictions in the multiverse is notoriously difficult. The problem is that an eternally inflating spacetime will produce an infinite number of all kinds of regions. So comparing the relative likelihood of one type versus another involves a comparison of infinite numbers. Cosmologists have attempted to regulate these infinities, in order to make sensible predictions, but in general the issue is still unresolved.

Questions

1. At the end of Chap. 9 we noted the remarkable fact that we live at a very special epoch when the density of matter is comparable to that of the vacuum: $\rho_m \sim \rho_v$. At much earlier times ρ_m was much greater and in the distant future it will be much smaller than ρ_v . Can you explain this fact using anthropic arguments? (For a complete explanation, you may need to know that the lifetime of a star like our Sun is about 10 Gyr.)
2. Which of the following parameters are life-changing and which are life-neutral: the electron charge, the neutrino mass, the strength of strong nuclear force, the large-scale curvature of the universe.
3. In some region of the multiverse the density of matter is measured to be the same as in our neighborhood, while the vacuum energy density is $\rho_v = 8\rho_{m0}$. At what redshift z_v did the vacuum dominated era begin in this region?
4. A population survey on some planet revealed that 5% of its territory is occupied by cities, 75% by rural areas, and the remaining 20% is not suitable for living. The population density (that is, the number of inhabitants per square kilometer) is 50 times higher in cities than in rural areas. What is the probability that a randomly picked inhabitant lives in a city? (This problem illustrates how volume fractions can be used to calculate probabilities.)
5. What ordering of integers would yield the fraction of odd integers equal to $2/3$? Is this ordering unique, or can you find other orderings that give the same answer?
6. Considering that the number of people who have ever lived is presently about 100 billion, use the doomsday argument to estimate the number of people who will ever live on Earth at 90% confidence level.
7. (a) Are you persuaded by the doomsday argument? If not, what do you think is wrong with it?
(b) Suppose humans will use bioengineering to evolve into some more advanced species within a few hundred years from now. Would the doomsday argument still apply?

22

Did the Universe Have a Beginning?

We have studied the early universe, its evolution, and its eternally inflating future. We are now poised to revisit a question that people have grappled with from the dawn of humanity: Did the universe have a beginning? Or has it always existed?

22.1 A Universe that Always Existed?

The notion that the universe has always existed is very appealing. It allows one to circumvent the avalanche of seemingly unassailable questions associated with the beginning of the universe. What caused the universe to appear? “Who/what” sets the initial conditions for the universe? Where did the “who/what” come from? This line of questioning is the endless regression that has haunted theologians, philosophers and scientists for millennia.

To address this issue, we need to investigate if it is possible to scientifically describe a universe that is eternal to the past as well as the future. Let’s begin by remembering that the steady state theory was of this sort. Observations favored the rival big bang, and the steady state theory was discarded. What if we consider an *oscillating* universe that undergoes a perpetual cycle of expansion and contraction—with a big bang followed by a big crunch followed by a big bang and so on? Such oscillating models were briefly considered in the 1930s, but soon they were found to be inconsistent with the second law of thermodynamics. The second law stipulates that each cycle of cosmic expansion is accompanied by an increase in entropy. If the universe had

undergone an infinite number of cycles in its past, the entropy would have reached its maximal value, and the universe would be in a state of thermal equilibrium. But we do not find ourselves in such a state. This is similar to the “heat death” problem of an eternal static universe that we mentioned in Sect. 5.1.1.

In 2002 Paul Steinhardt and Neil Turok introduced a new version of the oscillating model, which they called “the cyclic universe”. As in the older models, each cycle begins with an expanding fireball. As the fireball expands, it cools, galaxies form and then a period of vacuum domination ensues. Once vacuum domination sets in, the universe begins to expand exponentially. This exponential expansion is very slow—it takes about 10 billion years for the universe to double in size. After trillions of years, the expansion slows down and eventually stops and turns into contraction. When the collapse is complete, the universe bounces back to start yet another cycle. In this scenario the universe has eternally been undergoing a sequence of expansion and contraction, and there seems to be no need for a beginning.¹

But what about the problem of entropy that plagued the original oscillating models? In Steinhardt and Turok’s scenario, the amount of expansion in a cycle exceeds the amount of contraction, so the overall volume of the universe grows. The entropy of our observable region is now the same as the entropy of a similar region in the preceding cycle, but the entropy of the entire universe has increased—simply because the volume has increased. The growth of volume and entropy are both unbounded, and thus the state of maximum entropy is never reached—it does not exist.

Another option for a universe without a beginning is suggested by eternal inflation. Most cosmologists think that a period of cosmic inflation preceded the big bang. Inflation ended in our local region, setting off a local big bang, but continues elsewhere. This naturally raises the question: is it possible that inflation and subsequent big bang events have been occurring in spacetime for all of past eternity? Perhaps our ancestral chain of bubble universes goes back to the infinite past?

It turns out, however, that the idea of a past-eternal universe, either in cyclic or eternally inflating form, runs into a fatal obstruction—as we shall now discuss.

¹The cyclic model was introduced as an alternative to inflation, but it is far from being fully developed. To ensure a transition from expansion to contraction, the model requires a scalar field with a judiciously designed energy landscape. It also gives no satisfactory description for the bounce from the big crunch to the big bang. So, as of now, it remains a work in progress.

22.2 The BGV Theorem

In 2003, Arvind Borde of Long Island University, Alan Guth and Alex Vilenkin proved a theorem, which implies that even though inflation is eternal to the future, it cannot be eternal to the past and must have had some sort of a beginning. Their conclusions also apply to an oscillating model, which must have had a beginning too.

Borde, Guth and Vilenkin (BGV) investigated what an expanding universe would look like to imaginary observers who are recording their histories as they move through the universe under the influence of gravity and inertia. The observers are presumed to be indestructible, so if the universe had no beginning, the worldlines of all such observers should extend to the infinite past. However, BGV showed, under very plausible assumptions, that this is impossible.²

To see why, let us imagine that the entire universe is sprinkled with a “dust” of inertial observers who are all moving away from one another. The existence of such a class of observers can be taken as the definition of an expanding universe. We will call these observers “spectators”. Now, let us consider another observer, the *space traveler*, who has been moving relative to the spectators for all eternity. The space traveler also moves by inertia, with his spaceship engines shut off. As he passes the spectators, they register his velocity.

Since the spectators are flying apart, the space traveler’s velocity relative to each successive spectator will be smaller than his velocity relative to the preceding one. Suppose, for example, the space traveler passes the Earth at 100,000 km/s and is now headed towards a distant galaxy, about a billion light years away. Since that galaxy is moving away from us at 20,000 km/s, when the space traveler reaches it, the spectators there will see him moving at 80,000 km/s.

If the velocity of the space traveler relative to the spectators gets smaller and smaller into the future, then his velocity should get larger and larger as we follow his history into the past. In the limit, his velocity approaches the speed of light. The key insight in the BGV paper is that this limiting velocity is reached in a finite time by the space traveler’s clock. The reason is due to time dilation—remember moving clocks tick slower. As we go backward in time, the speed of the space traveler approaches that of light, and his clock

²The BGV theorem states that if the universe is, on average, expanding, then its history cannot be indefinitely continued into the past. The theorem allows for some periods of contraction, but on average expansion is assumed to prevail.

essentially freezes—from the spectator’s point of view. The space traveler himself does not notice anything unusual—his time flows from one moment to the next. Like the histories of the spectators, the space traveler’s history should extend into the infinite past.

The fact that the time elapsed by the space traveler’s clock is finite tells us that we do not have his full history. In technical language, physicists say that the space traveler’s world line is incomplete. This means that some part of the past history of the universe is missing; it is not included in the model. Thus, the assumption that the entire spacetime can be covered by an expanding dust of observers has led to a contradiction, and therefore it cannot be true.

The BGV theorem is very general. It makes no assumptions about the material content of the universe and does not even assume that gravity is described by Einstein’s equations. So, if Einstein’s theory requires some modification, the theorem would still hold. The only assumption it makes is that the universe is expanding at some non-zero rate (no matter how small). This should be satisfied by any model of eternal inflation. Thus we are led to the conclusion that past-eternal inflation without a beginning is impossible.³

Past-eternal cyclic models without a beginning are also ruled out. The volume of the universe increases in each cycle, hence the universe expands on average. This means that the space traveler’s velocity increases on average as we go back in time, and approaches the speed of light in the limit. Thus the same conclusions apply.

22.2.1 Where Does This Leave Us?

In Chap. 7 we discussed how scientists were drawn to the steady state theory over the big bang, because it avoided the question of a cosmic beginning. However, despite philosophical prejudice, the data had spoken and scientists had to press forward trying to uncover what they could about the universe in the context of the big bang model. Along the way they discovered inflation, and then eternal inflation. Our picture of the universe beginning with a one-time big bang event has given way to a much grander picture of an eternally inflating spacetime constantly spawning local big bangs. This worldview has the same spirit as the steady state theory, and many people once again hoped that maybe on a far greater scale the universe is indeed

³Note that it follows from the BGV theorem that the universe of the steady state model must also have a beginning.

eternal—with ancestor bubbles nucleating ad infinitum into the past. Now, however, we know that this is not possible. And once again, the beginning of the universe must be tackled head on.

22.2.2 A Proof of God?

Theologians and some religiously inclined scientists have often welcomed any evidence for the beginning of the universe, regarding it as evidence for the existence of God. On the other hand, a number of atheist scientists have argued that modern science leaves no room for God. A series of science-religion debates have been staged, with atheists like Richard Dawkins, Daniel Dennett and Lawrence Krauss combatting theists like William Lane Craig. The BGV theorem has often been raised as evidence for God by the theistic side.

It seems unlikely that science can disprove the existence of God, especially considering that “God” means different things to different people. Are we talking about the God of the Hebrew Bible or the rationalistic God of Spinoza and Einstein? A scientific proof of God based on the BGV theorem appears even more dubious.

The cosmological argument for the existence of God, dating back to Aquinas, consists of two parts. The first part is apparently very straightforward: “Everything that begins to exist has a cause. The universe began to exist. Therefore, the universe has a cause.” The second part affirms that the cause must be God. In the next chapter we will deconstruct this argument. We will argue that modern physics can describe the emergence of the universe as a physical process that does not require any supernatural cause.

Summary

The Borde–Guth–Vilenkin theorem says that the history of any expanding universe cannot be indefinitely continued into the past. An immediate implication is that inflation, even though it may be eternal to the future, cannot be eternal to the past and must have had a beginning.

We are thus faced with the question of what happened before inflation. And whatever the answer is, we can keep asking: “And what happened before that?” Thus the question of *how the universe began* is still enveloped in a cocoon of mystery.

Questions

1. Do you find the notion of an eternal universe to be preferable to a universe that somehow came into being from nothing?
2. Inflation is almost certainly eternal to the future. Is it eternal to the past too? Why/why not?
3. (a) State the BGV theorem. (b) Suppose future research shows that Einstein's theory of gravity needs to be modified. Could this invalidate the BGV theorem? (c) Does the BGV theorem make any assumption as to whether or not the universe is spatially finite or infinite?
4. Imagine a static closed universe, which existed in this state from past eternity until some moment when the static phase ended and inflation began. Does this model contradict the BGV theorem? If not, do you see any other problems with this scenario?
5. Do you think a scientific proof of a beginning implies there had to be a creator?

23

Creation of Universes from Nothing

If inflation is eternal, then the beginning of our local universe, about 14 billion years ago, was preceded by an unknown number of ancestor bubble universes. Although we do not know how far back the chain goes, we now believe that there had to be a beginning (as discussed in Chap. 22). So, how *did* it all begin? Eternal inflation pushes the ultimate beginning so far back into the past that we are unlikely to ever have direct observational evidence helping us to answer this question. Yet it must be addressed. It is arguably the most profound mystery that exists and it is at the core of our cosmological yearning. Here we will try to elucidate the speculative yet scientific attempts to explain how an embryonic seed universe emerged.

23.1 The Universe as a Quantum Fluctuation

We have already learned that the vacuum is a frenzied place filled with virtual particles and fields constantly fluctuating in and out of existence. Vacuum fluctuations live off borrowed energy for exceedingly small time intervals, in accordance with Heisenberg's uncertainty principle. For example, a spontaneously nucleated electron-positron pair will vanish in about a trillionth of a nanosecond. Heavier particle-antiparticle pairs live even briefer lives. If particles and antiparticles can spontaneously appear, why can't a fledgling universe?

This seemingly crazy idea was put forward by Edward Tryon, of the City University of New York, in the early 1970s.¹ Tryon suggested that the entire

¹At about the same time, a very similar idea was proposed by Piotr Fomin in the Soviet Union.



Fig. 23.1 The total charge in a closed universe is zero. Field lines emanating from a positive charge at the north pole will converge at the south pole—thus there must be an equal negative charge at the south pole

universe emerged as a quantum fluctuation out of the quantum vacuum. Not surprisingly, this idea was taken as a joke at first—there is a gaping difference between subatomic particles nucleating for about a trillionth of a nanosecond (or less) and a massive universe appearing and lingering for billions of years! Nonetheless, Tryon realized that there are no physical laws that forbid this happenstance. You might be thinking “But what about energy conservation?” Surely Lucretius was correct when he said: “Nothing can be created from nothing”. So how can a universe containing at least 10^{53} kg of matter suddenly appear? Here Tryon invoked a well-known fact: closed universes have zero energy. We emphasized several times earlier in this book that gravitational energy is negative. And it follows from general relativity that in a closed universe the negative energy of gravity exactly balances the positive energy of matter, so the total energy is zero. Another conserved quantity is electric charge, and once again it turns out that the total charge must vanish in a closed universe.

The latter statement is easy to understand using a two-dimensional analogy. Imagine a two-dimensional closed universe, which we can picture as the surface of a globe (see Fig. 23.1). Suppose we place a positive charge at the north pole of this universe. Then the lines of the electric field emanating from the charge will wrap around the sphere and converge at the south pole. This means that a negative charge of equal magnitude should be present there. Thus, we cannot add a positive charge to a closed universe without adding an equal negative charge at the same time. The total charge of a closed universe must therefore be equal to zero.

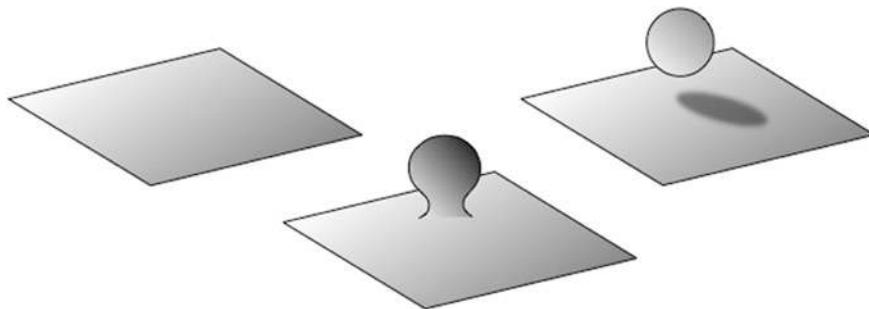


Fig. 23.2 Creation of a closed universe out of the vacuum

The creation of a closed universe out of the vacuum is illustrated in Fig. 23.2. A region of flat space begins to swell, taking the shape of a balloon. At the same time, a large amount of matter is spontaneously created in that region. The balloon eventually pinches off—becoming a closed universe, filled with matter, that is completely disconnected from the original space. Of course, it is very unlikely for such a huge quantum fluctuation to occur. But in quantum theory any process that is not strictly forbidden by conservation laws will happen with some probability. Also, since a nucleated closed universe does not borrow any energy from the quantum vacuum, it can persist for an indefinitely long time without violating the uncertainty principle.

A potential problem with Tryon's idea is that it is hard to understand why such a large universe would appear. It would be much more likely for a Planck sized universe to fluctuate out of the vacuum (as in the spacetime foam picture in Fig. 20.2). Even if we concede that observers require a certain amount of space to evolve, our universe still appears to be much larger than necessary to host observers.

Another, more fundamental issue with Tryon's scenario is that it does not really describe a universe appearing from nothing. The vacuum is what we call empty space. But as we know from Einstein's general relativity, even empty space can bend and warp, and have various geometries, such as the open, closed and flat models we have already encountered. Also, from quantum mechanics we know that the vacuum has energy density and tension, particles and fields. So the vacuum is very much “something”, which itself has to be presupposed to exist. As Alan Guth put it, “In this context, a proposal that the universe was created from empty space is no more fundamental than a proposal that the universe was spawned by a piece of rubber. It might be true, but one would still want to ask where the piece of rubber came from”. We shall now discuss how Tryon's idea can be extended, to describe quantum creation of the universe from “literally nothing”.

23.2 Quantum Tunneling from “Nothing”

Suppose we have a closed spherical universe, filled with a false vacuum and containing a certain amount of ordinary matter. Suppose also that this universe is momentarily at rest, neither expanding nor contracting. Its future will depend on its radius. If the radius is small, then matter is compressed to a high density, and its gravity will cause the universe to collapse. If the radius is large, the vacuum energy dominates and the universe will inflate. Small and large radii are separated by an energy barrier, which cannot be crossed unless the universe is given a large expansion velocity.

This is according to classical general relativity. But quantum physics provides another option: the universe can tunnel through the energy barrier, from a small to a large radius, and start inflating. The tunneling probability depends on the contents of the universe and on its radius. An interesting question is what happens as we let the radius become smaller and smaller. Remarkably, one finds that in the limit of vanishing radius there is still a well-defined, non-zero probability for the tunneling to occur. But a universe with a vanishing radius is no universe at all! One also finds that in this limit the universe that emerges after tunneling contains only vacuum energy and no matter.

Thus, one arrives at a mathematical description for a universe to be spontaneously created out of “nothing”. Here “nothing” means a state that contains no matter, and in addition it is also completely devoid of space and time. The “tunneling from nothing” picture was introduced by Vilenkin in 1982 and was later developed by Linde, Valery Rubakov, Alexei Starobinsky, and Yakov Zeldovich.

The tunneling proposal allows for the newborn universe to be filled with different types of vacua. As usual in quantum theory, we cannot tell which of these possibilities is actually realized, but can only calculate their probabilities. The mathematical analysis of the tunneling process shows that the initial radius of the universe right after tunneling is determined by the value of the vacuum energy density ρ_v :

$$R = c \left(\frac{3}{8\pi G \rho_v} \right)^{1/2} \quad (23.1)$$

High-energy vacua correspond to small radii. For example, if the universe emerges with a GUT scale vacuum energy density, its initial size would be $R \sim 10^{-28}$ cm. One also finds that the highest probability is obtained for

the universe having the largest vacuum energy and the smallest initial size.² Once the universe is formed, it immediately starts expanding, due to the repulsive gravity of the vacuum. This provides the beginning for the scenario of eternal inflation.

At this point you may be wondering: “What caused the universe to pop out of nothing?” Surprisingly, no cause is needed. If you have a radioactive atom, it will decay, and quantum mechanics gives the decay probability in a given interval of time. But if you ask why the atom decayed at this particular moment and not another, the answer is that there is no cause: the process is completely random. Similarly, no cause is needed for a quantum creation of the universe.

Another question you might be asking is: what happened before the tunneling? But we can’t meaningfully talk about time before the tunneling. As *St. Augustine* put it centuries ago: “The world was made not in time but simultaneously with time. There was no time before the world”. Time only has meaning if something is changing. Without space and matter time does not exist.

23.2.1 Euclidean Time

Quantum creation of universes is similar to quantum tunneling through energy barriers in ordinary quantum mechanics. An elegant mathematical description of this process can be given in terms of the so-called Euclidean time. This is not the kind of time you measure with your watch. It is expressed using imaginary numbers like $i = \sqrt{-1}$ and is introduced only for computational convenience. Making the time Euclidean has a peculiar effect on the character of spacetime: the distinction between time and the three spatial dimensions disappears, so instead of spacetime we have a four-dimensional space. This Euclidean-time description is very useful, as it provides a convenient way to determine the tunneling probability and the initial state of the universe as it emerges from the tunneling.

The birth of the universe can be graphically represented by the spacetime diagram in Fig. 23.3. Here we show one time and one spatial dimension. Time flows from the bottom of the figure upwards—it starts out being Euclidean, and then switches to regular time at the instant labeled

²In 1983 James Hartle and Stephen Hawking proposed an alternative model for a quantum description of the creation of the universe, called the no-boundary proposal. We will not discuss this model in detail, except to note that it gives opposite predictions to the tunneling-from-nothing proposal, assigning the highest probability to the smallest vacuum energy and largest initial size of the universe.

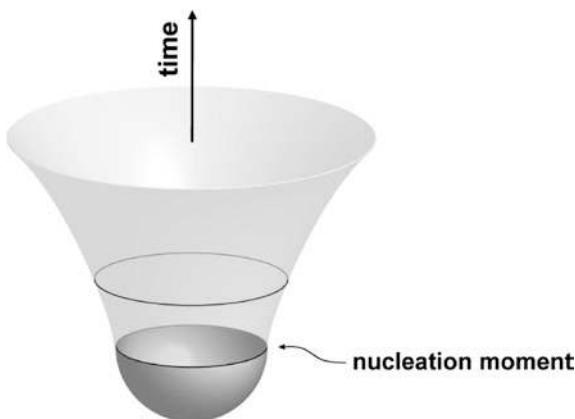


Fig. 23.3 Spacetime diagram of the universe tunneling from nothing. Two out of the three spatial dimensions are not shown. The *dark hemisphere* is the Euclidean region of spacetime, and the *circle* at its boundary represents the spherical universe at the moment of nucleation. The *larger circle* is the universe at a later moment of time

“nucleation moment”. The dark hemisphere represents quantum tunneling and the light surface above the nucleation moment is an inflating spacetime. A salient feature of this model is that there are no singularities. A Friedmann spacetime has a singular point of infinite density and curvature at the beginning, where the mathematics of Einstein’s equations breaks down (see Fig. 5.5). In contrast, the Euclidean spherical region has no such points; it has the same finite curvature everywhere. It thus gives a mathematically consistent description of how the universe could be born.

23.3 The Multiverse of Quantum Cosmology

As we already mentioned, in the tunneling from nothing proposal a universe can emerge with any one of a variety of values for the vacuum energy and its initial size. Also, although the nucleated universe must have a closed geometry, it need not be perfectly spherical. A range of shapes is allowed. Because of the quantum mechanical nature of the tunneling, we cannot determine which of these possibilities has been realized. All we can do is calculate the probability for a universe to emerge in one of the allowed states. This suggests that there should be a multitude of other universes, which started out differently from our own. We shall refer to this ensemble of universes as the “multiverse of quantum cosmology”.



Fig. 23.4 Multiple disconnected closed universes, each capable of producing an unbounded number of open bubble universes

You might imagine that closed universes pop out of nothing like bubbles in a glass of champagne, but this analogy would not be quite accurate. Bubbles pop out in the liquid, but in the case of universes, there is no space outside. Each universe in the quantum cosmology multiverse has its very own space and time, and is completely disconnected from all the others (see Fig. 23.4).

The most probable, and thus the most numerous, universes in the ensemble are the ones with the smallest initial radius and the largest false vacuum energy density. Our best guess, then, is that our own universe also originated in this way. The probability to nucleate larger universes decreases with size, and goes to zero in the limit of an infinite radius. Thus an infinite open universe has a precisely zero chance of nucleating from “nothing”, and all the universes in the ensemble are necessarily closed.

Whatever the initial vacuum state of a newborn universe, it will spawn an unlimited number of bubbles (or “bubble universes”) filled with other vacua. The entire vacuum landscape will be explored during the course of eternal inflation. Thus, each member of the quantum multiverse ensemble is a multiverse in its own right, including bubble universes of all possible kinds.

23.4 The Meaning of “Nothing”

We have described how an inflating seed could emerge from what seems to be literally nothing—a state with no space, no time, and no matter. However, the birth of the universe by quantum tunneling is described by the same laws

of physics that govern its subsequent evolution. So the laws must somehow be “there” prior to the universe. And the laws of physics are definitely not “nothing”. This is why we put the word “nothing” in quotation marks.

The notion of “nothing” transmuting into “something” under the spell of abstract laws of nature is deeply mystifying. If there is no time and space, where and how are these laws encoded? After all, the laws of physics have been carefully deduced over centuries by observing and experimenting with matter in space and time. They are supposed to describe our physical reality. Yet if the universe quantum tunneled as prescribed by the laws, then it seems that the laws must be more fundamental than the universe itself. One could become a “matheist” and assert that the laws of physics exist outside of space and time, much like a theist assigns the ultimate first cause to God. Or perhaps the fundamental laws and space and time emerged together?

We have stumbled far into the unknown. But we will press on with the optimistic hope that as the boundaries of scientific enquiry expand, what is currently unknowable might one day be known.

Summary

Quantum theory suggests that a small closed universe could be spontaneously created out of nothing. The newborn universe can materialize with a variety of sizes and can be filled with different types of vacua. The probabilistic nature of the tunneling suggests that an ensemble of universes can nucleate—we call this ensemble the multiverse of quantum cosmology. The most probable universes are the ones having the smallest initial size and the highest vacuum energy. Once such a universe emerges, it starts expanding rapidly because of the repulsive gravity of the vacuum. This provides a beginning for the scenario of eternal inflation.

No cause is needed for the quantum creation of a universe—it is a completely random process, like the decay of a radioactive atom.

Questions

1. What is the total energy of a closed universe? What is its total electric charge?
2. The lifetime of a virtual pair of particles Δt depends on the energy of the particles E . Higher energy particles have shorter lifetimes. Quantitatively, $E \cdot \Delta t \sim \hbar$, where \hbar is the reduced Planck constant. Use this relation to calculate approximately how long it takes for a virtual electron-positron pair to vanish. (The rest mass energy of one electron is 8.187×10^{-14} J and $\hbar = 1.055 \times 10^{-34}$ Js.)

3. Is there a bound to how long a spontaneously nucleated closed universe can live? Why/why not?
4. An important characteristic of elementary particles is their baryon number. This number is equal to 1 for nucleons and -1 for antinucleons. The baryon number is conserved in all particle interactions that have been studied so far. On the other hand, grand unified theories suggest that this conservation law is only approximate and must be violated in high-energy interactions. Assuming that the universe was created from nothing, can you tell whether the baryon conservation law is approximate or exact?
5. How does the quantum vacuum differ from a state of “nothing”?
6. Quantum tunneling from nothing allows a microscopic closed universe filled with false vacuum to pop out of nowhere and to immediately begin inflating. Is it more likely for such a universe to nucleate with a higher or lower false vacuum energy density? Does this make sense to you?
7. A universe which quantum tunnels from nothing must be closed. Does this mean that our local universe has a spherical geometry? Explain your answer. (Hint: See the discussion of bubble geometry in Sect. 18.4.)
8. Why does the tunneling from nothing proposal imply that there should be a multitude of other universes?
9. James Hartle and Stephen Hawking suggested an alternative description for the quantum origin of the universe. In their model, the most probable initial states have the lowest vacuum energy density and the largest radius.
 - (a) Do you find this picture intuitively plausible?
 - (b) Suppose the initial state of the universe was a large, empty, very low-energy universe, as the Hartle-Hawking model suggests. What would be the following evolution of this universe? (Hint: remember the possibility of bubble nucleation by tunneling “up”; see Sect. 18.3.) Is there any place where we could live in such a universe? Do you think this model can be ruled out observationally?
10. What do we mean by the following terms: observable universe, bubble universe, multiverse, and the multiverse of quantum cosmology?

24

The Big Picture

We are now going to summarize what we have learned about the universe, starting with things we are confident about and going up step by step in the level of conjecture, all the way to very speculative ideas. We shall discuss the “big picture” that has emerged and the answers it gives to the “big questions” of cosmology.

24.1 The Observable Universe

24.1.1 What Do We Know?

We can confidently trace our cosmic origins to a hot, dense fireball that erupted 13.7 billion years ago. This primeval fireball permeated all of the observable universe and was rapidly expanding. It consisted of matter and antimatter in almost equal amounts, with a slight excess of particles over antiparticles. The expansion caused the fireball to cool and set in motion a series of major cosmological events.

As the temperature dropped, particles annihilated with their antiparticles, and at about one second after the big bang essentially all the antiparticles had vanished.¹ At that time the fireball was a mix of photons, neutrinos, electrons, neutrons, protons, and dark matter particles. The photons outnumbered matter particles by about a billion to one. Within the first few

¹To be precise, antineutrinos are still present in the universe today, in about the same amounts as neutrinos. They did not annihilate because they interact so weakly.

minutes the universe had cooled enough for the first atomic nuclei to form. A long uneventful period followed, and then at about 400,000 years electrons and nuclei combined into electrically neutral atoms. Consequently, the universe became transparent to light, so photons were free to stream ahead unimpeded. These photons now come to us from all directions in the sky as the cosmic background radiation.

Minute differences in the cosmic radiation intensity tell us that some of the photons came from regions that were slightly more or less dense. These tiny density fluctuations in the early universe grew larger and larger as gravitational attraction dialed up the contrast between high and low density regions. In about a billion years these fluctuations had turned into the first galaxies. Dark matter played a crucial role in this process. Galaxies continued to grow via mergers. Larger structures like clusters and superclusters also continued to emerge until about five billion years ago, when the universe became dominated by the vacuum energy and the structure formation process came to a halt.

We now know that atomic matter makes up only about 5% of the total energy (or mass) density of the universe; dark matter, which has yet to be directly observed, contributes about 26%, and the vacuum energy density about 69%. The contribution to the energy density from photons and neutrinos has been redshifted to insignificance today.

This account of cosmic history is supported by a wealth of observational data, and there is very little doubt that it is basically correct. Most physicists would agree with the sentiment of Yacov Zel'dovich when he proclaimed: “I am as sure of the big bang as I am that the Earth goes around the Sun.” And yet, the big bang theory is not completely satisfactory. The theory postulates that the initial fireball has rather peculiar properties: it must be very hot, expanding, space-filling, uniform, and it must have tiny fluctuations. Furthermore, the big bang theory does not say where the fireball came from in the first place. These issues are addressed in the theory of cosmic inflation.

24.1.2 Cosmic Inflation

Inflation is a hypothetical period of fast, accelerated expansion in the early cosmic history. It is driven by the repulsive gravity of a high-energy false vacuum. At the end of inflation, the false vacuum decays, igniting a hot fireball of particles and radiation. As we discussed in Chap. 16, a period of inflation naturally explains the otherwise mysterious features of the fireball, which had to be postulated before. Once the fireball ignites, the universe evolves along the lines of the standard big bang cosmology. Thus, the end of inflation plays the role of the big bang in this theory.

The details of cosmic inflation depend on the energy landscape of particle physics, of which little is known. At present, therefore, inflation is not a specific model, but rather a framework encompassing a wide class of models with different energy landscapes. But despite the variety of models, some observational predictions of inflation are very robust and have been convincingly confirmed by the data. By now inflation has become the leading cosmological paradigm. With increasing accuracy of observations we can expect to learn more about the inflationary epoch in the early universe and about particle physics at very high energies.

24.2 The Multiverse

24.2.1 Bubble Universes

A remarkable property of cosmic inflation is its eternal character. The end of inflation is triggered by quantum, probabilistic processes and does not occur everywhere at once. In our cosmic neighborhood inflation ended 13.7 billion years ago, but it still continues in remote parts of the universe, and other “normal” regions like ours are constantly being formed.² The new regions appear as tiny, microscopic bubbles (or “islands”) and immediately start to grow. They keep growing without bound, but the gaps between them are also growing. Thus there is always room for more bubbles to form, and their number proliferates *ad infinitum*. This never-ending process is what we call eternal inflation. We live in one of the bubbles and can observe only a small part of it, so for all practical purposes we live in a self-contained bubble universe.

If the energy landscape of particle physics includes a multitude of vacuum states, then bubbles filled with each type of vacuum have a nonzero probability to form. So inevitably, an unlimited number of bubbles of all possible types will be formed during the course of eternal inflation. The constants of nature, such as the masses of elementary particles, Newton’s gravitational constant, etc., will take a variety of different values in different bubble types.

This multiverse picture explains the long-standing mystery of why the constants of nature appear to be fine-tuned for the emergence of life. The reason is that intelligent observers exist only in those rare bubbles in which, by pure chance, the constants happen to be just right for life to emerge. The rest of the multiverse remains barren, but no one is there to complain about that.

²Inflation is not eternal in some special energy landscapes, but such models appear to be rather artificial.

24.2.2 Other Disconnected Spacetimes

Even though inflation is eternal to the future, it must have had a beginning in the past. An attractive proposal for the beginning, which avoids the endless succession of questions “And what happened before that?” is the idea that spatially closed inflating universes can quantum-mechanically nucleate out of “nothing”.

The newly born universes can have a variety of different shapes and sizes and can be filled with different types of vacua. As usual in quantum theory, we cannot tell which of these possibilities is realized in any given universe, but can only calculate their probabilities. One finds that the most probable initial states are those of small size and high vacuum energy density.

All nucleated universes are completely disconnected from one another. They start from different initial states, but in the process of eternal inflation get populated by bubbles of all possible types.³ At later times these universes “forget” their initial state and become statistically very similar to one another.

24.2.3 Levels of the Multiverse

The term “multiverse” has different meanings depending on the context in which it is used. In this book we talked about the multiverses of eternal inflation and of quantum cosmology. In fact, one can distinguish at least three levels of the multiverse.⁴

Level 1: an individual bubble universe. When viewed from inside, a bubble universe is spatially infinite and includes an infinite number of horizon regions like ours (see Chap. 18). In this sense, a bubble universe is a multiverse in its own right. Different horizon regions in a bubble have very similar physical properties, but differ in detail: the pattern of galaxy distributions and the forms of life will differ from one region to another. As Max Tegmark lightly sums it up, physicists in different horizon regions will take the same physics classes, but they would study different things in history.

Level 2: the multitude of bubble universes. This is the multiverse of eternal inflation, populated with bubbles of all possible types. In this kind of multiverse, observers in different types of bubbles would learn different things not only in history class, but in physics class too.

³The initial nucleation process can be regarded as the state of “nothing” branching into embryo universes with different initial conditions. This is similar to the many worlds interpretation of quantum mechanics, where each of the spawned universes will further branch into a multitude of parallel universes.

⁴Note that our classification of multiverse levels is somewhat different from that of Max Tegmark in his book “Our Mathematical Universe”.

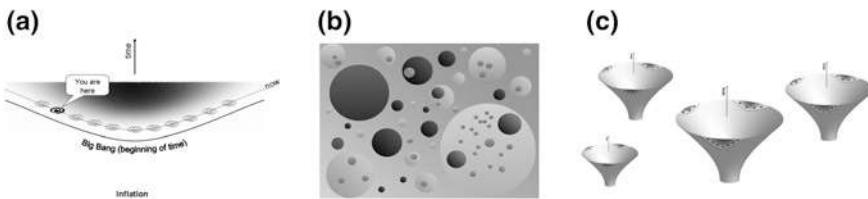


Fig. 24.1 Three levels of the multiverse. **a** Level 1: our entire observable universe is most likely a very small domain in one bubble universe. An infinite number of regions like ours may exist beyond our horizon. **b** Level 2: the eternally inflating multiverse consists of an ensemble of bubbles formed during eternal inflation. **c** Level 3: multiple disconnected spacetimes produced by quantum tunneling from nothing. This ensemble of disconnected spacetimes is the “multiverse of quantum cosmology”

Level 3: the multiverse of quantum cosmology. The members of this multiverse are all level-2 multiverses. They differ mostly by their initial states and become statistically very similar at later times.

The three levels of the multiverse are illustrated in Fig. 24.1. Each higher level involves a higher level of speculation. The existence of other horizon regions beyond our own is rather uncontroversial. The observable universe is nearly uniform, so it is natural to expect that the uniform distribution of galaxies and radiation extends beyond our horizon. Even if inflation is not eternal, there should be a multitude of horizon regions, and in this sense a level 1 multiverse would still exist.

The level 2 multiverse assumes the existence of multiple bubble types, with different physical properties, populated by the mechanism of eternal inflation. The successful prediction of the cosmological constant provides indirect evidence for this kind of multiverse. Direct evidence may also be found by looking for signatures of collisions of our bubble with other bubbles and for “failed bubbles” in the form of black holes with a special distribution of masses.

We may never be able to test the existence of the level 3 multiverse observationally. Nevertheless, it may be necessary for the completeness and consistency of our worldview.

24.2.4 The Mathematical Multiverse and Ockham’s Razor

The Standard Model of particle physics, combined with general relativity, can be expressed as a single mathematical equation, shown in Fig. 24.2. This equation encodes the physical character of our bubble universe. In an eternally inflating multiverse, the physical character of bubble universes can vary greatly from one bubble to another, so the equations expressing *local* physical

quantum mechanics spacetime gravity

$$W = \int_{k < \Lambda} [Dg][DA][D\psi][D\Phi] \exp \left\{ i \int d^4x \sqrt{-g} \left[\frac{m_p^2}{2} R - \frac{1}{4} F_{\mu\nu}^a F^{a\mu\nu} + i\bar{\psi}^i \gamma^\mu D_\mu \psi^i + (\bar{\psi}_L^i V_{ij} \Phi \psi_R^j + \text{h.c.}) - |D_\mu \Phi|^2 - V(\Phi) \right] \right\}$$

other forces matter Higgs

Fig. 24.2 Physics/mathematics underlying our everyday life. This “picture” is a single equation that has the power to explain almost everything about our physical world. It embodies quantum mechanics, general relativity and the Standard Model of particle physics. This shorthand mathematics encodes a great deal of information and readily turns into pages and pages of equations when physicists use it for particular problems. It is consistent with almost every experiment we have ever performed on Earth. Note We still do not have a unified theory of gravity and quantum mechanics, and it does not account for neutrino masses and dark matter. Credit From S. Carroll, *The Big Picture: On the Origins of Life, Meaning, and the Universe Itself* (Dutton 2016)

laws will also vary. However, underlying this diversity, physicists expect to find a single fundamental theory of nature—an equation or a set of equations describing the entire multiverse. The same equations should also describe the nucleation of inflating multiverses from nothing. These could be the equations of string theory, or of some yet undiscovered fundamental theory. Finding this “theory of everything” is the holy grail of theoretical physics. But whatever it is, the question naturally comes to mind: Why this theory? Who or what selected it from the infinity of all possible mathematical laws?

Max Tegmark, now at the Massachusetts Institute of Technology, suggested a possible answer, which is as simple as it is radical: there is a physical universe corresponding to each and every mathematical structure.⁵ This amounts to adding another level to the multiverse.

Level 4: the multiverse of mathematical structures. This includes, for example, a Newtonian universe governed by the laws of classical mechanics, without quantum theory and relativity, as well as a “universe” described by the natural set of numbers 1, 2, 3, ..., supplemented by the rules of addition and multiplication, or by a 5-dimensional cube. The mathematical structures in some of the universes are intricate enough to allow the emergence of self-aware substructures, like you and us. Such universes might be rare, but of course they are the only ones that can be observed. All universes in this

⁵Tegmark actually goes further than that and asserts that the universe is a mathematical structure.



Fig. 24.3 Max Tegmark's mathematical multiverse includes universes described by all possible mathematical laws. Apart from some of his more outlandish ideas, Tegmark has made important contributions to the study of the CMB and the large-scale distribution of galaxies. At present his research is focused on the physics of the brain. Photo courtesy Max Tegmark

mathematical multiverse are completely disconnected from one another, but Tegmark asserts that they are all equally real (Fig. 24.3).

Such a dramatic extension of physical reality may raise an objection that it runs against one of the most cherished principles of science—"Ockham's razor". As formulated by the 14th century English philosopher William of Ockham, this principle states "Entities should not be multiplied beyond necessity." In our context this can be interpreted to say that if you can explain all observations with both a simple theory and a more complex theory, then the simple theory is preferable. Then why add an infinite ensemble of mathematical structures if our observations can be accounted for by a single equation as in Fig. 24.2? Similar objections can be raised against level 1, 2 and 3 multiverses: they appear to be very wasteful of space, matter, and universes.

But in mathematics it is often simpler to describe an entire ensemble than to specify individual members of that ensemble. For example, the natural set of integers 1, 2, 3, ... can be generated by a short computer program, while specifying a single large integer may require many more bits of information.

Similarly, it is much simpler to write down the equations governing eternal inflation than it is to specify all the precise initial conditions that would be needed to describe one particular bubble universe. The process of eternal inflation is simple and economical; it effortlessly generates great diversity with very little input, much like cell-division and natural selection. In this sense, it is the opposite of “wasteful” and seems to be in line with Ockham’s razor. As Alan Guth predicts, “Given the plausibility of eternal inflation, I believe that soon any cosmological theory that does not lead to the eternal reproduction of universes will be considered as unimaginable as a species of bacteria that cannot reproduce.”

Tegmark’s mathematical multiverse idea is not without its problems. There is an infinite number of mathematical structures. If all of them are equally probable, then the probability of any one of them is exactly zero. (Note that, in contrast, the number of possible bubble types is finite, even though it may be extremely large.) Moreover, the number of mathematical structures proliferates with increasing complexity. By the principle of mediocrity, we should then expect to find ourselves living in a horrendously large and cumbersome mathematical structure. But we tend to find that nature prefers simplicity.

As of now, the mathematical multiverse idea does not have much following in the physics community. The level 2 multiverse of eternal inflation was very controversial in the 1980s and 90s, but the tide has gradually changed, and it is now widely accepted by mainstream cosmologists (with some notable exceptions). The level 3 multiverse of quantum cosmology is regarded as an intriguing possibility, but progress in this area should await the development of the theory of quantum gravity.

24.3 Answers to the “Big Questions”

We opened this book asking a few basic questions: Is the universe finite or infinite? Has it existed forever? If not, when and how did it come into being? Will it ever end? Philosophers and theologians have been arguing about these questions for millennia, and one might expect that all possible answers have already been anticipated. However, the worldview suggested by modern cosmology is not what anyone expected. It does not give “yes” or “no” answers to the big questions, but rather maintains that each of the opposing options has some element of truth.

Each bubble universe is infinite when viewed from the inside. And yet the entire inflating multiverse is spatially closed and finite. Due to the non-trivial spacetime geometry, this multiverse can contain an unlimited number of bubble universes.

The multiverse of eternal inflation is eternal only to the future: it had a beginning, but will have no end. It began as a tiny closed space by popping spontaneously out of “nothing”. Our local region, however, will come to an end. It will end in a big crunch, after being engulfed by a bubble of negative energy density.

24.4 Our Place in the Universe

Ever since Copernicus, advances in science have been casting doubt on the naïve view that our planet and our civilization play some special role in the cosmos. Not only did we discover that the Earth is not the center of the Solar System, but the Sun itself is an unremarkable star at the outskirts of a rather typical galaxy. Now, according to the new worldview, our bubble universe contains an infinite number of galaxies, which harbor an infinity of earths identical to our own. Confronted with this worldview, it is hard to feel special! What meaning can then we assign to our lives and our civilization?

The universe is governed by mathematical laws, which make no reference to meaning. The concept of meaning is created by humans, and it is up to us to give meaning to the universe. We may be insignificant on the grand cosmic scale, but we could extend our significance well beyond our planet Earth. We don't know how widespread intelligent life is, but the chances are that we are the only intelligent civilization in the entire observable region. If so, we are custodians of a huge chunk of real estate, 80 billion light years across. We have now reached the point where we can either self-destruct or colonize our galaxy and beyond. This makes us nothing but significant. Our crossing the threshold to a space-colonizing civilization would be an enormous breakthrough. It would make the difference between us being a “flicker” civilization that blinks in and out of existence, and being a civilization that spreads through much of the observable universe, and possibly transforms it.

Appendix A

This appendix covers some mathematical details of relativistic cosmology, and is suitable for readers who are familiar with calculus.

The Friedmann Equation

In Chap. 7 we introduced the concept of a scale factor $a(t)$, which is defined as the factor by which distances between comoving objects in a homogeneous and isotropic universe at cosmic time t differ from their distances at the present time t_0 . Thus, by definition, $a(t_0) = 1$. We also showed that the expansion rate of the universe, as characterized by the Hubble parameter, is related to the scale factor as:

$$H = \frac{\dot{a}(t)}{a(t)}, \quad (\text{A.1})$$

where an overdot denotes the rate of change, or the time derivative, $\dot{a} = \frac{da}{dt}$.

The magnitude of H is related to the energy density by the Friedmann equation, which we shall now derive.

Let us consider a comoving spherical region of radius $R = a(t)R_0$, as we did in Sect. 8.1. The total energy of a test particle that lies on the boundary of the sphere is given by (see Eq. 8.1)

$$\frac{1}{2}mv^2 - \frac{GMm}{R} = \frac{1}{2}mR^2 \left(H^2 - \frac{8\pi}{3}G\rho \right). \quad (\text{A.2})$$

Here, $v = HR$ is the particle's velocity, $M = \frac{4\pi}{3}R^3\rho$ is the mass of matter enclosed by the sphere, and ρ is the mass density. We shall assume that the universe is dominated by matter, so the force due to pressure is negligible; then M does not change during the course of expansion.

Since energy is conserved, the magnitude of the right-hand side of Eq. (A.2) at any time is the same as it is at present; hence we can write

$$R^2 \left(H^2 - \frac{8\pi}{3}G\rho \right) = R_0^2 \left(H_0^2 - \frac{8\pi}{3}G\rho_0 \right),$$

where zero subscripts indicate quantities evaluated at the present time. Dividing by R^2 (where $R = a(t)R_0$) and factoring out H_0^2 on the right-hand side, we can rewrite this equation as

$$H^2 - \frac{8\pi G\rho}{3} = \frac{H_0^2}{a^2}(1 - \Omega_0). \quad (\text{A.3})$$

Here, $\Omega_0 = \frac{\rho_0}{\rho_{c0}}$ is the present value of the density parameter $\Omega = \rho/\rho_c$, where

$$\rho_c = \frac{3H^2}{8\pi G} \quad (\text{A.4})$$

is the critical density.

Equation (A.3) is the famous Friedmann equation, which is a statement of energy conservation. Note that although we derived it here assuming that pressure is negligible (which is not so at early and late times), it is valid in general. Another useful form of the Friedmann equation is obtained by multiplying Eq. (A.3) with a^2 . This gives

$$\dot{a}^2 = \frac{8\pi G\rho}{3}a^2 + H_0^2(1 - \Omega_0). \quad (\text{A.5})$$

Solutions in Different Cosmic Epochs

The observed density of the universe is very close to the critical density. Hence, to a good approximation we can set $\Omega_0 = 1$. Then the Friedmann equation simplifies to

$$\dot{a}^2 = \frac{8\pi G}{3}\rho a^2. \quad (\text{A.6})$$

In general, the mass density ρ includes contributions from matter (atomic and dark matter), radiation, and the vacuum; thus

$$\rho = \rho_m + \rho_r + \rho_v \quad (\text{A.7})$$

The three contributions have different a -dependence,¹

$$\rho_m = \frac{\rho_{m0}}{a^3}, \quad \rho_r = \frac{\rho_{r0}}{a^4}, \quad \rho_v = \text{const}, \quad (\text{A.8})$$

and come to dominate the universe at different epochs, as we discussed in Sect. 11.7. Their present magnitudes are

$$\begin{aligned} \rho_{m0} &= 2.7 \times 10^{-27} \text{ kg/m}^3, & \rho_{r0} &= 7.9 \times 10^{-31} \text{ kg/m}^3, \\ \rho_v &= 6.0 \times 10^{-27} \text{ kg/m}^3 \end{aligned}$$

We can use Eq. (A.8) to estimate the redshift z_{eq} at the end of the radiation era, when radiation and matter densities are equal:

$$\rho_m(t_{eq}) = \rho_r(t_{eq}) \Rightarrow \frac{\rho_{m0}}{a^3(t_{eq})} = \frac{\rho_{r0}}{a^4(t_{eq})}. \quad (\text{A.9})$$

Simplifying,

$$\frac{\rho_{m0}}{\rho_{r0}} = \frac{1}{a(t_{eq})} = z_{eq} + 1, \quad (\text{A.10})$$

where we have used $\frac{1}{a(t)} = z + 1$ (this is Eq. 7.8). Using the measured values of ρ_{m0} and ρ_{r0} , we find $z_{eq} \approx 3400$.

We can also use Eq. (A.8) to calculate the redshift z_v at the end of the matter era, when the vacuum energy density begins to dominate, $\rho_v = \rho_m$. Using Eqs. (A.8) and (7.8), we can write

$$\rho_v = \rho_{m0}(1 + z_v)^3 \quad (\text{A.11})$$

Solving this for z_v we obtain $z_v = 0.30$.

¹ Neutrinos are nearly massless particles, and their density scales in the same way as the density of photons, $\rho \propto a^{-4}$. The neutrino density is therefore included in the radiation density ρ_r .

Radiation Era

When the universe is dominated by radiation, the Friedmann equation (A.6) takes the form

$$\dot{a}^2 = \frac{8\pi G\rho_{r0}}{3a^2} \quad (\text{A.12})$$

where we have used $\rho \approx \rho_r$ with ρ_r from Eq. (A.8). Taking the square root, we have

$$a \frac{da}{dt} = \left(\frac{8\pi G\rho_{r0}}{3} \right)^{1/2} \quad (\text{A.13})$$

This has the solution

$$a(t) = C_r t^{1/2} \quad (\text{A.14})$$

where the constant coefficient C_r is given by

$$C_r = \left(\frac{32\pi G\rho_{r0}}{3} \right)^{1/4} \quad (\text{A.15})$$

Substituting this solution in Eq. (A.1), we find the Hubble parameter

$$H = \frac{1}{2t} \quad (\text{A.16})$$

The density can now be found from

$$\rho = \frac{3H^2}{8\pi G} = \frac{3}{32\pi Gt^2} \quad (\text{A.17})$$

The temperature at time t is

$$T = \frac{T_0}{a(t)} = \frac{T_0}{C_r t^{1/2}} \quad (\text{A.18})$$

where $T_0 = 2.7\text{ K}$ is the present CMB temperature (the first part of Eq. (A.18) is Eq. 11.8).

You can now use Eqs. (A.17) and (A.18) to calculate the mass density and the temperature of the universe at any time during the radiation era. If t is in seconds, then

$$\rho = \frac{4.5 \times 10^8}{t^2} \text{ kg/m}^3 \quad (\text{A.19})$$

$$T = \frac{10^{10}}{t^{1/2}} \text{ K} \quad (\text{A.20})$$

These equations were introduced without derivation in Chap. 14. (In deriving the last relation, we substituted ρ_{r0} into Eq. (A.15) to find $C_r \approx 2 \times 10^{-10}$).

We can approximate the time of matter-radiation equality t_{eq} by substituting Eqs. (A.14) and (A.15) into Eq. (A.10) to get

$$t_{eq} = \left(\frac{\rho_{r0}}{\rho_{m0}} \right)^2 \frac{1}{C_r^2} \approx 68,000 \text{ yrs.} \quad (\text{A.21})$$

This estimate is not very accurate because our expression for $a(t)$ was found by assuming the radiation density is much bigger than the matter density (and all other densities). But at the time of equality t_{eq} , the two densities are equal, so the actual behaviour of $a(t)$ is more complex during the “cross-over” period. A more accurate numerical calculation gives $t_{eq} \approx 51,000$ yrs.

Matter Era

When the universe is dominated by matter, the Friedmann equation is

$$\dot{a}^2 = \frac{8\pi G \rho_{m0}}{3a} \quad (\text{A.22})$$

Following the same steps as above, we find the solution

$$a(t) = C_m t^{2/3} \quad (\text{A.23})$$

where

$$C_m = (6\pi G \rho_{m0})^{1/3} \quad (\text{A.24})$$

The Hubble parameter and the mass density are now given by

$$H = \frac{2}{3t} \quad (\text{A.25})$$

and

$$\rho = \frac{1}{6\pi G t^2} \quad (\text{A.26})$$

Substituting the solutions Eqs. (A.23) and (A.24) into Eq. (A.11) we can approximate the time of vacuum domination $t_v \approx 11.5$ Gyr. Once again, this estimate is not very precise because our expression for $a(t)$ is not accurate near t_v . A more detailed calculation gives $t_v \approx 10$ Gyr.

Note: Here we have defined vacuum domination to start at the time when the energy density of matter becomes equal to the energy density of the vacuum. However, we showed (in question 14 of Chap. 9) that the condition for accelerated expansion to begin is that $\rho_v > \rho_m/2$. This means that accelerated expansion actually begins sooner than the time t_v that we calculated here. This is why we have mentioned several times in the book that acceleration began about 5 billion years ago.

Vacuum Dominated Era

During the vacuum dominated phase, which only began recently, the energy density is given by $\rho_v = \text{const}$, and the Friedmann equation becomes

$$\dot{a}^2 = \frac{8\pi G \rho_v}{3} a^2 \quad (\text{A.27})$$

Taking a square root, we have

$$\dot{a} = H_v a \quad (\text{A.28})$$

where

$$H_v = \sqrt{\frac{8\pi G \rho_v}{3}} \quad (\text{A.29})$$

The solution is

$$a(t) = C_v e^{H_v t} \quad (\text{A.30})$$

where $e \approx 2.72$ is the base of natural logarithms, and $C_v = \text{const.}$

Inflation

During inflation the universe is also dominated by vacuum energy. So the Friedmann equation is the same as (A.27), except the inflationary vacuum energy density ρ_v is much greater than it is at present. The scale factor is still given by (A.30). This tells us that in a time interval Δt the universe expands by a factor $e^{H_v \Delta t}$.

In Chap. 16 we defined the doubling time t_D as the time it takes the universe to double in size. We can find this from

$$e^{H_v t_D} = 2 \quad (\text{A.31})$$

which gives

$$t_D = H_v^{-1} \ln 2 = 0.69 H_v^{-1} \quad (\text{A.32})$$

Flatness Problem

In a universe filled with ordinary matter or radiation, the density ρ is rapidly driven away from the critical value ρ_c , so in order to have $\Omega = \frac{\rho}{\rho_c} \approx 1$ at present, the universe must have started with Ω extremely close to 1 at some early time. We discussed this fact, known as the flatness problem, in Chap. 15; now we will show how it follows from the Friedmann equation Eq. (A.3).

Dividing both sides of Eq. (A.3) by H^2 and using the definition of the critical density Eq. (A.4), we have

$$1 - \Omega = \frac{H_0^2}{H^2 a^2} (1 - \Omega_0). \quad (\text{A.33})$$

Now, from Eq. (A.1), $Ha = \dot{a}$, and thus

$$1 - \Omega \propto \frac{1}{\dot{a}^2}, \quad (\text{A.34})$$

where \propto means “proportional to.”

In a universe filled with ordinary matter or radiation, the speed of expansion \dot{a} decreases with time, due to the attractive gravitational force. Then it follows from Eq. (A.34) that $|1 - \Omega|$ grows with time. In other words, the universe deviates more and more from the critical density. Using the solutions (A.14) and (A.23), we find that $(1 - \Omega) \propto t$ in the radiation era, and $(1 - \Omega) \propto t^{2/3}$ in the matter era. For example, from cosmic time $t = 1\text{s}$ till the end of the radiation era at $t_{eq} \approx 2 \times 10^{12}\text{s}$, $(1 - \Omega)$ grew by a factor of 2×10^{12} , and from t_{eq} to the present time t_0 it grew approximately by a factor $(\frac{t_0}{t_{eq}})^{\frac{2}{3}} \approx 2 \times 10^3$. Overall, $(1 - \Omega)$ increased by a factor 4×10^{15} between $t = 1\text{s}$ and now.² This means that for $|1 - \Omega| < 0.1$ now, we must fine-tune $|1 - \Omega|$ to be less than 2×10^{-17} at $t = 1\text{s}$.

Equation (A.34) also explains how the flatness problem is solved by cosmic inflation. During inflation the expansion of the universe accelerates, so \dot{a} grows and $|1 - \Omega|$ decreases with time. From (A.30), $\dot{a} \propto e^{H_v t}$ and

$$(1 - \Omega) \propto e^{-2H_v t} \propto a^{-2}. \quad (\text{A.35})$$

This shows that the density approaches the critical density exponentially fast. If, for example, inflation expanded the universe by a factor of 10^{50} , then Ω was driven closer to 1 by a factor of 10^{100} .

Note: By the definition of the doubling time n , (here we denote the doubling time by “n” to relate to Fig. 16.4), we can write the scale factor as

$$a(t) \propto 2^n \quad (\text{A.36})$$

allowing us to recast Eq. (A.35) in terms of the doubling time as,

$$(1 - \Omega) \propto 2^{-2n} \quad (\text{A.37})$$

² Here we disregard the relatively small change in $(1 - \Omega)$ during the vacuum dominated era, which started only recently.

Further Reading

Here is a list of books that may be helpful for further exploration of some of the topics that we have covered. We have grouped them according to their main focus, although most of these books cover other topics as well. In the last group we included some books which are critical of the ideas of inflation, string theory, and the multiverse. The “Further viewing” list includes some excellent web courses on various aspects of cosmology.

Relativity and Quantum Physics

Deutsch, David. *The Fabric of Reality*. New York: Viking Adult, 1997.

Einstein, Albert. *The Meaning of Relativity (5th edition)*. Princeton University Press, 2004.

Greene, Brian. *The Fabric of the Cosmos*. New York: Knopf, 2004.

Thorne, Kip S. *Black Holes and Time Warps: Einstein's outrageous legacy*. New York: W. W. Norton & Company, 1995.

Unification of Forces

Greene, Brian. *The Elegant Universe: Superstrings, Hidden Dimensions, and the Quest for the Ultimate Theory*. New York: W. W. Norton and Company, 1999.

- Randall, Lisa. *Warped Passages: Unraveling the Mysteries of the Universe's Hidden Dimensions*. Harper Perennial, 2006.
- Weinberg, Steven. *Dreams of a Final Theory: The Scientist's Search for the Ultimate Laws of Nature*. Pantheon, 1992.
- Wilczek, Frank. *The Lightness of Being: Mass, Ether, and the Unification of Forces*. Basic Books, 2008.
- Zee, Anthony. *Fearful Symmetry: The search for Beauty in Modern Physics*. Macmillan, 1986.

Big Bang Cosmology

- Coles, Peter. *Cosmology: A Very Short Introduction*. Oxford University Press, 2001.
- Harrison, Edward. *Cosmology: The Science of the Universe*. Cambridge University Press, 2000.
- Kirshner, Robert P. *The Extravagant Universe: Exploding Stars, Dark Energy, and the Accelerating Cosmos*. Princeton University Press, 2002.
- Livio, Mario. *The Accelerating Universe: Infinite Expansion, the Cosmological Constant, and the Beauty of the Cosmos*. Wiley, 2000.
- Rees, Martin. *Just Six Numbers: The Deep Forces That Shape the Universe*. Basic Books, 2001.
- Silk, Joseph. *The Big Bang*. W. H Freeman & Co., 1988.
- Weinberg, Steven. *The First Three Minutes: A Modern View of the Origin of the Universe*. New York: Basic Books, 1993.

Cosmic Inflation

- Guth, Alan. *The inflationary Universe*. New York: Perseus Books Group, 1997.

The Multiverse

- Davies, Paul. *Cosmic Jackpot*.
- Greene, Brian. *The Hidden Reality*. New York: Knopf, 2011.
- Kaku, Michio. *Parallel Worlds*.
- Susskind, Leonard. *The Cosmic Landscape: String Theory and the Illusion of Intelligent Design*. New York: Little, Brown and Company, 2005.
- Vilenkin, Alex. *Many Worlds in One: The Search for Other Universes*. New York Hill and Wang, 2006.

Quantum Origin of the Universe

- Krauss, Lawrence. *A Universe from Nothing: Why there is something rather than nothing*. New York: Free Press 2012.
- Hawking, Stephen. *A Brief History of Time* (10th anniversary edition). Bantam, 1998.

Life in the Universe

- Davies, Paul. *The Eerie Silence: Renewing Our Search for Alien Intelligence* (2nd edition). Houghton Mifflin, 2010.
- Gribbin, John. *Alone in the Universe: Why Our Planet is Unique*. Wiley, 2011.

The Big Picture

- Carroll, Sean. *The Big Picture: On the Origins of Life, Meaning, and the Universe Itself*. Dutton, 2016.
- Hawking, Stephen and Mlodinov, Leonard. *The Grand Design*. Random House Publishing Group, 2012.
- Tegmark, Max. *Our Mathematical Universe: My Quest for the Ultimate Nature of Reality*. New York: Knopf, 2014.

Alternative Views on Inflation, String Theory and the Multiverse

- Smolin, Lee. *The Trouble with Physics: The Rise of String Theory, The Fall of a Science, and What Comes Next*. Houghton Mifflin Harcourt, 2007.
- Steinhardt, Paul and Turok, Neil. *Endless Universe*. Doubleday (5th or Later Edition), 2007.

Further viewing

- Whittle, Mark. *Cosmology: The History and Nature of Our Universe*. Virginia: The Teaching Company, 2008.
- Carroll, Sean. *Dark Matter, Dark Energy: The Dark Side of the Universe*. Virginia: The Teaching Company, 2007.
- Alex Filippenko. *Understanding the Universe: An Introduction to Astronomy*, 2nd Edition. Virginia, The Teaching Company, 2007.

Index

A

Absolute future and past 52
Absolute zero 157
Acceleration 7, 9, 15, 16, 20, 21, 32, 82, 358
Accelerator .*See* particle accelerator 202
Action-at-a-distance 27
Alchemy 14
Alpha centauri 51
Alpha decay 150
Alpha particle 147, 149
Alpher, Ralph 7
Andromeda galaxy 6, 103
Anthropic principle 307
Anthropic selection 301
Antibaryon 220
Antimatter 222
Antineutrinos 302
Antiparticles 202–204, 211, 220, 221
Antiquarks 211
Aristarchus 4
Aristotle 2, 60
Atomic nuclei 8
Atoms 2, 6, 8, 39, 98, 115, 120, 144, 147
Augustine, Saint 337
Axiom 63, 64, 66, 69, 81

B

Baryogenesis 220
Baryon number 220, 221
Baryons 220
Bell curves 313, 315
Bell Telephone Laboratories 163
Bentley, Richard 26, 97
BICEP 262
Big bang 6, 7, 9, 91, 115, 117–119
Big bang nucleosynthesis 189, 191, 192, 198
Big crunch 92, 125
Biology 165
Black body radiation 160
Black holes 5, 77, 79
Blueshift 166
Bohr, Niels 150
Boltzmann constant 88
Boltzmann, Ludwig 88
Bolyai, Janos 67
Bondi, Herman 114, 192
Boomerang detector 262
Borde-Guth-Vilenkin theorem (BGV) 329
Bosons 205, 207, 208, 211
Bottom quark 206
Bousso, Raphael 296

- Brahe, Tycho 5
 Branes 295
 Brightness 7, 27, 71, 103, 105
 Brout, Robert 207
 Bubble nucleation 271, 274, 284, 286, 296, 297
 Bubble universe
 collisions 275, 284, 285
 contracting 279, 285
 expanding 276, 279, 284–286
- C**
- Carbon 98
 Carter, Brandon 309, 322
 Cepheid stars 104, 106
 Chaotic inflation 181, 306
 Charge, conservation of 35
 Charmed quark 206
 Chibisov, Gennady 247
 Circular motion 4, 21
 Civilizations 322, 323
 Classical mechanics 144
 Closed-universe model 93
 COBE satellite 7
 Coincidence problem 10
 Collapse of the wave function 149
 Compact dimensions 294
 Compactification 294, 296
 Computer simulations 183, 218, 273
 Confessions
 Augustine 337
 Confinement, of quarks 211
 Conservation laws 202, 281, 335
 Constants of nature 294, 296, 298
 Copenhagen interpretation 151, 152
 Copernicus, Nicolaus 4
 Cosmic Background Explorer (COBE)
 satellite 7
 Cosmic background radiation 166, 233
 Cosmic clones 279
 Cosmic distance ladder 103, 109
 Cosmic egg 1
 Cosmic horizon 118, 120
 Cosmic strings 217, 218
 Cosmological constant 84, 94
 Cosmological constant problem 304
 Cosmological principle 114
 Creation of the universe 335
 Creation stories 1
 Critical mass density 1
 Curvature of space 72, 84
 Curvature of spacetime 80
 Cyclic universe 238
- D**
- Dark energy 1, 9, 10, 119
 Dark matter 1, 2, 4–6, 10
 Dark night sky paradox 27
 Democritus 2
 Density parameter 1, 6, 9, 10, 128
 Density perturbations 256
 De Sitter spacetime 119
 De Sitter, Willem 113
 Deuterium 188, 190
 Dicke, Robert 163, 165
 Dimensions, extra 294
 Disorder 87, 88
 Distance determination 101
 DNA 197
 Domain walls 215–218
 Doomsday argument 321–323
 Doppler shift 3
 Doubling time 359, 360
 Down quark 206
 Dwarf galaxies 104
- E**
- Earth 1–5, 16–20, 22, 23, 29, 31, 32, 34, 44, 45, 51, 56–58, 61, 68, 74, 77, 81, 92, 97, 98, 106, 109, 117–120
 Eddington, Arthur 75
 Edwin Hubble 107
 Efstatihou, George 318

- Einstein, Albert 31
 Electric charges 35, 39
 Electric forces 35
 Electromagnetic radiation 36, 39
 Electromagnetic spectrum 38, 79
 Electromagnetic waves 36, 38, 39, 79,
 98, 121, 143, 144, 152
 Electromagnetism 35, 39, 144
 Electron 8, 10, 144, 146, 148, 149, 152
 Electron volt 202
 Electroweak force 211, 214, 215
 Element abundances 187
 Elementary particles 161, 206, 220,
 237, 345
 Elements 1, 7, 63, 98, 114
 Elliptical galaxies 74
 Empty space, gravity of 2
 End of the 32
 Energy 22, 23, 25, 28, 74, 85, 103,
 126, 144, 358
 Energy conservation, law of 25
 Energy landscape 295
 Englert, Francois 207
 Entropy 87
 Epicurus 2, 3
 Epicycles 3, 4
 Equilibrium, thermal 88
 Erwin Schrodinger 150
 Eternal inflation 269, 274
 Ether 33, 34
 Euclidean geometry 48, 64–66, 71, 72,
 92
 Euclidean time 337
 Euclid of Alexandria 63, 65
 European Center for Nuclear
 Research(CERN) 203
 Event horizon 78, 119
 Everett, Hugh, III 151
 Everett interpretation of quantum
 mechanics 151
 Evolution, Darwin's theory of 321
 Expansion of the universe
 accelerating 137, 138
 decelerating 137
 exponential 239
- F
- Fermi, Enrico 324
 Fermions 205, 206, 305
 Field
 electric and magnetic 35, 36, 38, 121
 energy density diagrams 250, 251
 gravitational 61, 74, 78, 256
 higgs 207, 208, 211
 inflaton 246, 247
 scalar 208, 237, 265, 282
 Final theory of nature 291
 Fine-tuning 304, 305, 307, 310
 Fireball,cosmic 6, 158, 161
 Five-nucleon gap 192
 Flatness problem 359
 Flat universe 92, 129
 Fluctuations
 density 256
 quantum 146, 247, 261, 277
 thermal 88, 210
 Fluxes 295
 Foam, spacetime 306
 Force 13–15, 19, 33, 74, 80, 105, 354
 Fractals 270
 Frequency 36, 38
 Friedmann, Alexander 7, 89, 90, 94
 Friedmann universe 90, 92, 129
 Fundamental theory of nature .*See* final
 theory 291, 298, 348
- G
- Galaxies 1, 2, 5–7, 22, 83, 86, 89, 93,
 106, 109, 111, 112, 114, 118,
 120, 121
 Galaxy clusters 4, 5
 Galileo 20, 31, 35, 60, 61, 72
 Gamma ray 39
 Gamow, George 7
 Gases, expansion of 98

- Gauge bosons 205, 207
 Gauss, Carl Friedrich 66
 General theory of relativity .*See* relativity theory, general 7, 28
 Genetic code 165, 197, 321
 Geodesic line 68
 Geometry 10, 64, 67, 73, 89, 130
 Giant galaxies 301
 Gluons 207, 211
 God 5, 40, 151
 Gold, Thomas 114
 Google 46
 Graceful exit problem 244, 245
 Grand unified theories 214, 215, 220, 222
 Grassmann numbers 305
 Gravitational collapse 247, 307
 Gravitational constant 18
 Gravitational energy 251, 334
 Gravitational field 56, 61, 74, 77, 78, 80
 Gravitational lensing 4, 5, 10, 75, 76, 81
 Gravitational waves 78, 81
 Gravitational waves, primordial 261
 Gravitons 206, 292
 Gravity 5, 7, 8, 16–20, 22, 26–28, 53, 60, 61, 72, 74, 77, 79, 80, 84, 85, 91, 92, 94, 105, 117–119, 125, 126, 129
 Great Wall 178
 Greeks, ancient 3, 63
 Guth, Alan 228, 235, 236, 244, 255, 260, 265
- H**
 Hartle, James 337
 Hawking, Stephen 337
 Heat-death problem 87
 Heinrich Wilhelm Olbers 28
 Heisenberg, Werner 145, 146
 Helium 7, 98
 Helmholz, Hermann von 87
 Herman, Robert 7
 Hierarchical clustering 185
 Higgs boson 205, 207, 211, 301
 Higgs field 207–210, 214, 216, 217, 237, 246, 282, 295
 Higgs particle 207
 Higgs, Peter 207
 High redshift supernova team 112
 Histories 272, 273, 280
 Homogeneous 70, 83, 84, 89, 92, 112
 Horizon, cosmic .*See* cosmic horizon 118, 120
 Horizon distance 119
 Horizon problem 229, 235, 241
 Hoyle, Fred 114
 Hubble constant 116, 127
 Hubble, Edwin 93, 111
 Hubble’s law 117, 118, 119, 121, 127
 Hubble time 118
 Hubble ultra deep field 181
 Hydrogen 3, 7, 98, 108, 149
 Hyperbolic geometry 6, 67, 70, 71, 92
- |
- Imaginary numbers 337
 Inertial motion 15, 60
 Inertial observer 31, 33, 39, 50, 54, 59
 Inflation 359, 360
 Inflaton field 246, 247
 Inhomogeneities, cosmic 184, 277
 Instability, gravitational 27, 179, 182
 Intelligent life 198, 275, 303, 305, 324, 351
 Inverse square law 18, 21, 71, 80
 Ionized 172
 Iron 193, 194, 210
 Irregular galaxies 176
 Isaac Newton 97
 Island universes 105, 106
 Isotope 187, 188
 Isotropic 70, 83, 84, 89, 92, 112

J

Johannes Kepler 5

K

Kelvin scale 157

Kepler, Johannes 21, 35

kicks, quantum 248, 249, 271, 272

Kinetic energy 22–25, 47, 126, 127

Kirshner, Robert 112

L

Landau, Lev 7

Large Hadron Collider (LHC) 203

Laser Interferometer Gravitational Wave Observatory (LIGO) 79

Last scattering 162, 229, 258

Leavitt, Henrietta Swan 103, 104, 106

Lemaitre, Georges 7, 93, 112

Length contraction 40, 44, 45, 46, 55

Length, units of 218

Leptons 206

Life, evolution of 197, 280, 305, 308, 316, 324

Light

frequency 36, 38

speed of 32–46

wavelength 36, 38, 98, 101, 159

Light cone, past and future 53, 120

Light years 3, 50, 51, 57, 83, 85, 100, 102, 103, 105, 106, 112, 120

Linde, Andrei 245, 246

Lithium 192, 193

Lobachevsky, Nikolai 67, 70

Local group of galaxies 175

Local supercluster 176

Lorentz factor 43, 44, 47

Lucretius 3

Luminosity 2, 103, 105, 109

M

Magnetic charge 220

Magnetic field 35, 36, 38, 121

Magnetic monopoles 220, 222, 232

Many worlds interpretation 152

Martel, Hugo 318

Mass density 1, 7, 10, 19, 74, 89, 127, 354, 358

Mass-energy equivalence 40, 46

Mathematical multiverse 347, 349, 350

Mather, John 167

Matter-antimatter asymmetry 221

Matter era 355, 360

Matter-radiation equality 357

Maxwell, James Clerk 35, 36, 39, 144

Maxwell's equations 35

Measure problem 319, 320

Mediocrity, principle of 313, 314

Mesons 221

Michelson-Morley experiment 35

Microwaves, cosmic 6

Milky way galaxy 6, 83

Minkowski, Hermann 48

Monopole problem 232, 242

Moon 1, 13, 17, 18, 20, 25, 32

Multiverse 281, 285, 286, 296, 308, 309, 315, 319, 320

Multiverse levels 346

Multiverse of quantum cosmology 338

Mukhanov, Viatcheslav 247

Muons 45

N

Nanosecond 333, 334

Nebulae 105–107, 109, 110

Negative pressure 85

Neutrinos 158, 189, 201, 206, 208, 212, 302, 343

Neutrons 10, 147

Neutron stars 2, 5, 79

New inflationary theory 251

Newton, Isaac 5, 13, 14, 31, 35

Newton's law of motion 13, 28

Newton's law of universal gravitation 13

Night sky paradox 27

- Nobel prize 7, 8, 58, 104, 111
 No-boundary proposal 337
 Nuclear physics 98
 Nuclear reactions 5, 7, 46
 Nucleation 244, 276
 Nucleons 157, 169, 191, 192, 212
 Nucleosynthesis
 big bang 189
 stellar 193
- O**
 Ockham's razor 347, 349, 350
 Olbers, Heinrich Wilhelm 28
 Olbers' paradox 27, 28
 Open universe 93, 259, 339
 Orders of magnitude 216, 293, 305, 307
 O-regions 279–281
 Oxygen 193, 196
- P**
 Pair annihilation 203, 204
 Parallax 101–103, 106
 Parallel universes 151
 Particle accelerator 202, 203
 Particle physics 9, 46, 152
 Penzias, Arno 6
 Perfect cosmological principle 114
 Periodic table 7
 Period-luminosity relation 104, 106
 Perlmutter, Saul 6, 8
 Phase transition 212
 Photons 39, 143, 144, 152
 Planck energy 215
 Planck length 293, 294, 306
 Planck, Max 36
 Planck satellite 168, 256, 257, 265
 Planck's constant 143
 Planetary system formation 194
 Planets 2–5, 13, 21, 24, 59, 74, 97, 98,
 110, 120, 144, 146
 Plasma 257, 258, 259, 261, 262
 Plato 2

- Pocket universes (see island universes) 105
 Polarization 261, 263
 Polchinski, Joseph 296
 Positrons 204, 213, 302
 Potential energy 22–25, 126, 237, 245,
 251, 271, 295
 Potential energy density 209, 237, 245
 Precession of Mercury's orbit 75, 80
 Primeval fireball 158, 161, 162, 343
 Principia 5, 13, 17, 21, 27
 Principle of mediocrity 313
 Principle of relativity 31, 39, 42, 54,
 59, 80
 Probabilities, in the multiverse 320
 Probabilities, quantum mechanics 346
 Proton decay 220
 Protons 10, 127, 147
 Ptolemy 3, 4
 Pythagoras 2
 Pythagorean Theorem 49, 50, 54
- Q**
 Quantization 152
 Quantum chromodynamics 214
 Quantum discreteness 144
 Quantum electrodynamics 39
 Quantum fluctuations 146, 152
 Quantum indeterminism 151
 Quantum kicks 248, 249, 271, 272
 Quantum mechanics 40, 143, 151, 152
 Quantum theory .See quantum mechanics 144, 148, 152
 Quantum tunneling 146, 147
 Quarks 157, 206
 Quasars 183, 184
- R**
 Radiation
 cosmic 163, 165, 169, 219, 344
 electromagnetic 36, 39, 158, 159
 Radiation era 355, 360

- Radioactivity 147
 Radio astronomy 7
 Radio waves 3, 115
 Random walk 271
 Recombination 161, 162, 168, 213, 256, 257
 Red giant 193
 Redshift 7, 8, 101, 109, 116, 117, 355
 Relativity of simultaneity 40, 41
 Relativity theory
 general 7, 28, 53, 73, 80, 83, 90
 special 39, 53, 59
 Repulsive gravity 9, 10
 Rest energy 47
 Riess, Adam 8, 112
 Rotation curve 2–4, 6, 10
 Rubakov, Valery 336
 Rubin, Vera 4, 6
 Rutherford, Ernest 7
- S**
- Sakharov, Andrei 221
 Scalar 208, 237, 265, 282
 Scalar fields 208, 237, 265, 282
 Scale factor 7, 9, 115–117, 128, 359, 360
 Scale-invariant 183, 248, 256, 260, 265
 Schmidt, Brian 6, 8
 Schrodinger, Erwin 147, 148, 150
 Schrodinger's cat 150
 Schwarzschild, Karl 78
 Schwarzschild radius 77
 Shapley, Harlow 105
 Simultaneity 41, 44
 Singularities 91
 Slipher, Vesto 110
 Slow roll inflation 245
 Smoot, George 167
 Solar System 2, 25, 31, 88, 105, 120
 Sound waves 99
 Space
 curvature of 69, 72
 empty 2, 39, 84, 105, 236, 335
 Spacetime 48, 50, 61, 77, 91
 Spacetime diagram 47, 52, 53, 119
 Spacetime event 47
 Spacetime interval 50, 57
 Special theory of relativity .*See* relativity theory, special 39, 40, 59
 Spectral lines 7, 106, 109
 Spectroscopy 98, 101
 Spectrum, electromagnetic 38, 79
 Spectrum of density fluctuations 248, 249
 Spin 194, 205, 292
 Spinoza, Baruch 40
 Spiral galaxies 2
 Spontaneous symmetry breaking 208
 Standard candles 7, 103, 106, 109
 Standard model of particle physics 205
 Starobinsky, Alexei 260, 335
 Stars 3, 4, 6, 10, 22, 46, 74, 75, 79, 87, 97, 101, 107, 114, 125
 Steady state cosmology 115
 Steinhardt, Paul 245, 274, 297
 Stellar nucleosynthesis 193
 Strange quarks 206
 String theory 361
 String theory landscape 296, 298
 Strong nuclear force 188, 206, 214, 215, 291, 296
 Structure formation 175
 Structure problem 232
 Sun 3, 17, 25, 144
 Super cluster 83, 167, 175, 176, 180, 232, 344
 Supernova Cosmology Project 194
 Supernovae 104, 105
 Surface of last scattering 162, 229, 258
 Susskind, Leonard 297
 Symmetry 70, 210, 211, 214, 215, 217
- T**
- Tau particle 206
 Tegmark, Max 346, 348, 349

- Temperature 5, 87
 Temperature anisotropies 256, 257, 265
 Tension 21, 85
 Thales 2
 Theory of everything .*See* final theory 348
 Thermal equilibrium 87, 88
 Thermal fluctuations 88
 Thermodynamics, second law of 87, 94
 Time 7, 110, 112, 113, 115, 117, 148
 Time dilation
 in special relativity 53
 gravitational 77
 Top quark 206
 Tritium 188, 190
 True vacuum 237, 243–246, 264, 270, 273, 278
 True vacuum bubbles 244
 Tryon, Edward 333, 334
 Tunneling from nothing 336
 Tunneling, quantum .*See* quantum tunneling 243, 282, 296, 336, 338, 339, 347
 Twin paradox 78
- U**
 Uncertainty principle 145
 Uncertainty, quantum 280
 Uniform circular motion 21
 Universe
 age of 118
 beginning of 115
 expanding 120
 infinite 27
 observable region of 88
 static 27
 structure of 89
 Up quark 206
- V**
 Vacuum 84, 85, 180, 208, 216, 236, 237, 240, 243, 250, 274, 296, 317
 Vacuum decay 243
 Vacuum defects 215
 Vacuum energy density .*See also* dark energy, cosmological constant 7
 Vacuum era 170, 171, 180
 Vacuum fluctuation 88
 Vector 15
 Velocity 2, 7, 127, 145
 Vilenkin, Alex 271, 273
 Virgo cluster 176, 177
 Virtual particles 304, 306
 Volume 19, 70, 71, 90
- W**
 Wave function 148–150
 Wavelengths
 defined 36
 W bosons 304
 Weak nuclear force 206–208, 211, 215
 Weinberg, Steven 318, 320
 White dwarf 2, 5, 105
 Wilkinson David 163
 Wilkinson Microwave Anisotropy Probe (WMAP) satellite 7
 Wilson, Robert 6, 7
 WMAP satellite .*See* Wilkinson Microwave Anisotropy Probe (WMAP) satellite 7
 Worldline 51, 52, 56, 57, 61, 86, 91, 119, 120
- X**
 X-rays 38
- Z**
 Z bosons 207, 208
 Zel'dovich, Yakov 344
 Zwicky, Fritz 4, 5